

# HOW TO MAKE USEFUL ONTOLOGIES FOR BIOMEDICINE

## SECTION I: OVERVIEW OF CURRENT APPLICATIONS OF ONTOLOGIES IN BIOINFORMATICS

**Goal:** *In this section, we will review current applications of ontologies in bioinformatics, with the aim of getting an implicit feel for what ontologies are and what they are used for. For each broad category of application, we will briefly describe the ontology, describe the use and then generalize that use to a broader application.*

Modern biomedical research is data-intensive and researchers seek tools to enable discovery in massive online databases. The e-science era has brought a proliferation in both data and databases, as well as an exponential growth in published literature[1]. Researchers must assimilate and integrate a growing amount of diverse information to do their work. This is clearly a challenge, and researchers are looking to computers to help them manage the information explosion[2]. In particular, they have turned to ontologies to structure their complex domain and to relate the myriad of data to shared understandings of biomedicine.

Ontologies in biomedicine span a spectrum, ranging from simple thesauri of term lists to highly expressive sources of biomedical knowledge. At the extreme of simplicity, the class names within bio-ontologies can be used as a controlled terminology for labelling data and reducing ambiguity in communication experimental results. At the other extreme, highly expressive ontologies are built containing detailed biomedical knowledge, enabling developers to create powerful computer reasoning applications that can make biomedical inferences on a massive scale. The knowledge represented in these ontologies also varies, ranging from information models and how to how to store and exchange data to encyclopaedic reference ontologies of biomedical knowledge and declarations of biomedical theory. Researchers need to be aware what domains of biomedicine are adopting ontologies and how ontologies can enable scientific work.

There are two complementary perspectives of bio-ontologies: a **content-oriented view**, concerned with the *specific ontologies* being created in biomedicine, and b) a **functional view**, dealing with *how* ontologies can be used to enable a diversity of biomedical applications. The content-oriented view has been well addressed in prior reviews [3-5], describing the activities of individuals and communities engaged in creating and improving ontologies, projects to accumulate and catalogue ontologies, and efforts to critique ontologies and to develop best practices for how ontology should be created. The functional view addresses how ontologies can be used, and can assist biomedical researcher understand the relevance of ontology to their work, as well as provide specific ideas for applications. In this review, we summarize biomedical ontology from the functional perspective, organizing the presentation according to how ontologies are used.

The range of ontology content and structure reflects the diversity of applications from the functional perspective—that ontologies are enabling a variety of types of biomedical use cases, serving as the outline for our review. Specifically, ontologies are being used in the following ways:

1. Reference for naming things
2. Representation of encyclopedic knowledge
3. Specification of information models
4. Specification of data exchange formats
5. Representation of semantics of data for information integration
6. Computer reasoning with data

We will select an example ontology to illustrate each of these use cases, expanding on each by providing:

- A general description of the ontology
- A possible application to motivate the use of the ontology
- Generalizations that may be derived based on the features of the ontology in that use case.

## 1) REFERENCE FOR NAMING THINGS: THE GENE ONTOLOGY

The requirement of “naming things” refers to the necessity of establishing a set of controlled terms for labeling entities in databases and datasets. This is perhaps the commonest task in biomedicine to enable computers to help researchers make sense of massive online datasets and carry out their analyses. The language of biomedicine contains many synonymous terms, abbreviations, and acronyms that can refer to the same thing. For example, the process of creating glucose is referred to using a variety of synonymous terms, including “glucose synthesis,” “glucose biosynthesis,” “glucose formation,” “glucose anabolism,” and “gluconeogenesis.”

It is challenging to unify diverse data sets in a consistent way when they describe similar entities that are labelled differently in different resources. An ontology provides a single name (the class name) for each entity it contains (though it can represent alternative names for that entity through the appropriate relations). The ontology can thus be used as a controlled terminology to label biomedical entities (genes, diseases, findings, etc) in a consistent way. In addition, the ontology can be augmented with terminological knowledge such as synonymy, abbreviations, and acronyms. Ontologies used in this manner enable the community to create integrated resources more easily and to contribute new terminological knowledge as the content of scientific discourse evolves.

### GENERAL DESCRIPTION: GENE ONTOLOGY

The Gene Ontology (GO) [6] is perhaps the canonical example of an ontology created for the primary purpose of providing controlled terms for naming things. The Gene Ontology Consortium developed GO in recognition of the fact that different Model Organism Database (MODs) describe the same functions, biological processes, and cell components of gene products using different terms. In order for the MODs to describe gene products in an unambiguous manner, the Gene Ontology Consortium was established to create a standard set of names of biological entities (GO terms) and their relationships. The GO consists of three ontologies, containing entities for naming biological processes, molecular functions, and cellular components of gene products (*Figure 1*). The three ontologies provide the terms to describe what the gene products do, where and when they act, and why they perform these activities.

- all : all ( 265317 )
    - GO:0008150 : biological\_process ( 180873 )
    - GO:0005575 : cellular\_component ( 159672 )
      - GO:0005623 : cell ( 118896 )
        - GO:0045177 : apical part of cell ( 320 )
        - GO:0043190 : ATP-binding cassette (ABC) transporter complex ( 162 )
        - GO:0045178 : basal part of cell ( 39 )
        - GO:0005933 : bud ( 305 )
        - GO:0000267 : cell fraction ( 2217 )
        - GO:0042995 : cell projection ( 1248 )
        - GO:0030428 : cell septum ( 57 )
        - GO:0043025 : cell soma ( 55 )
        - GO:0009986 : cell surface ( 923 )
        - GO:0051286 : cell tip ( 70 )
        - GO:0030312 : external encapsulating structure ( 816 )
        - GO:0031026 : glutamate synthase complex ( 5 )
        - GO:0042763 : immature spore ( 44 )
        - GO:0005622 : intracellular ( 90178 )
          - GO:0031255 : lateral part of motile cell ( 0 )
        - GO:0031252 : leading edge ( 272 )
        - GO:0016020 : membrane ( 35794 )
          - GO:0030496 : midbody ( 23 )
        - GO:0042597 : periplasmic space ( 186 )
          - GO:0001917 : photoreceptor inner segment ( 4 )
        - GO:0030427 : site of polarized growth ( 368 )
        - GO:0031254 : trailing edge ( 12 )
          - GO:0031317 : tripartite ATP-independent periplasmic transporter complex ( 3 )
      - GO:0008372 : cellular component unknown ( 33752 )
      - GO:0031012 : extracellular matrix ( 1231 )
      - GO:0005576 : extracellular region ( 10411 )

*Figure 1 - The Gene Ontology: The Gene Ontology as displayed in the Amigo Browser (amigo.geneontology.org). The Cellular Component branch of GO is expanded, showing that it comprises a hierarchically-organized set of terms describing the components making up cells, with children being related to parent terms via is-a relations. The GO terms are used to provide a controlled terminology for annotating biomedical databases and for creating computable biomedical assertions.*

The entities (represented by the terms) in the GO have *is-a* and *part-of* relations to other entities, providing the basis for representing biological knowledge. While these relations are not necessary to exploit the value of GO as a controlled terminology for names, they do enable computer reasoning applications, which can recognize that and can infer subsumption or composition by tracing the *is-a* and *part-of* relations, respectively.{OBOL}

## APPLICATION

The Gene Ontology has enabled all of the Model Organism Databases (MODs) to assert the functions, processes, cellular components associated with gene products in an unambiguous manner. To make these assertions, a list of GO terms is associated with each gene product in a process referred to as “annotation”<sup>1</sup> (Figure 2).

---

<sup>1</sup> The annotations created by the Model Organism Databases represent assertions about biology that hold the potential to be true for all individuals.

In this study, we report the isolation and molecular characterization of the *B. napus* PERK1 cDNA, that is predicted to encode a novel receptor-like kinase. We have shown that like other plant RLKs, the kinase domain of PERK1 has serine/threonine kinase activity. In addition, the location of a PERK1-GTP fusion protein to the plasma membrane supports the prediction that PERK1 is an integral membrane protein...these kinases have been implicated in early stages of wound response...

**Function:** protein serine/threonine kinase activity ; GO:0004674 (IDA)  
**Component:** integral to plasma membrane ; GO:0005887 (IDA)  
**Process:** response to wounding ; GO:0009611 (NAS)

*Figure 2- Ontology-based Annotation: An example of use of Gene Ontology to create assertion annotations based on biomedical text. In the excerpt from biomedical text shown, the molecular function (serine/threonine kinase activity), cellular component (integral membrane protein), and biological process (wound response) of PERK1, are summarized using the appropriate GO terms from each of the three GO ontologies.*

Creating such ontology-based annotations is highly valuable for both querying databases as well as analyzing high throughput data:

- **Simple queries:** researchers can search GO annotations to find all genes that have particular involved in particular biological processes, having certain molecular functions, or located in a specific cellular component.
- **GO based analysis of high throughput data:** These analyses use GO codes for gleaning biomedical insights into experimental results, and generally include the following tasks:
  - **Find over-represented GO categories** in a list of genes: If a group of genes have similar experimental results (such as sharing the same cluster in a microarray data set), the researcher can look at the GO codes associated with the genes in that cluster to establish common “biological themes” shared by that group of genes.[7]
  - **Binning:** obtain a broad view of the distribution of major GO terms in a list of genes by combining similar (but more granular) GO terms.[8]
  - **Clustering Genes** on GO terms: group together functionally related genes based on GO terms. Knowledge in the Gene Ontology guides the analysis of GO terms to perform this grouping. Gene clusters are determined by calculating an annotation-based distance between genes, taking into account all GO terms that are common to the pair and terms which are specific to each gene. The gene clusters are usually displayed using a dendrogram or a graph, based on a matrix containing the inter-gene distances.[9]

#### GENERALIZATIONS TO OTHER ONTOLOGIES AND APPLICATIONS

The use of ontologies to provide a controlled terminology for naming things has broad applications in biomedicine. In fact, many other projects are currently using ontologies precisely for this purpose:

1. **Indexing the biomedical literature:** The terms in ontologies can be used to index the literature for improving search as well as enabling text processing applications. The Medical Subject Headings (MeSH) is a terminology created by the National Library of Medicine for indexing the medical literature [10]. MeSH provides a standard set of terms that medical librarians use to describe the main topics covered in

papers, the species studies, funding source, and other attributes. Originally conceived to help librarians carry out literature search, the standard names provided by MeSH have proven useful for augmenting natural language processing methods for text processing, extraction, and classification [11-14]. The use of MeSH for providing names for biomedical entities in these applications is analogous to how names in the Gene Ontology used as annotations enable analysis of large microarray data sets: articles sharing MeSH terms have similar content, and gene products sharing GO codes have similar function, are involved in similar biological processes and are located at similar cellular locations.

**2. Integration and Access to Cancer Data:** Standard set of terms from an ontology are being used by large databases for indexing data and linking them to other resources. Cancer research and clinical practice require tight integration of large amounts of molecular and clinical data. The NCI Thesaurus, being developed by the National Cancer Institute, is designed to integrate molecular and clinical cancer-related information [15]. It provides a controlled terminology that enables researchers to integrate, retrieve, and relate diverse data collected in cancer research. It covers topics such as cancers, findings, drugs, therapies, anatomy, genes, pathways, cellular and subcellular processes, proteins, and experimental organisms. In addition, the NCI Thesaurus represents how the entities relate to each other in a description logic framework that enables curators to maintain the integrity and extend the informational power of the terminology. The terminology enables scientists to label experimental results in a standard manner, as well as to link their research findings to disease and molecular patterns [16]. Associating research data with ontology terms also enables efficient search and retrieval of that data, by querying different levels within the ontology [17] (*Figure 3*).

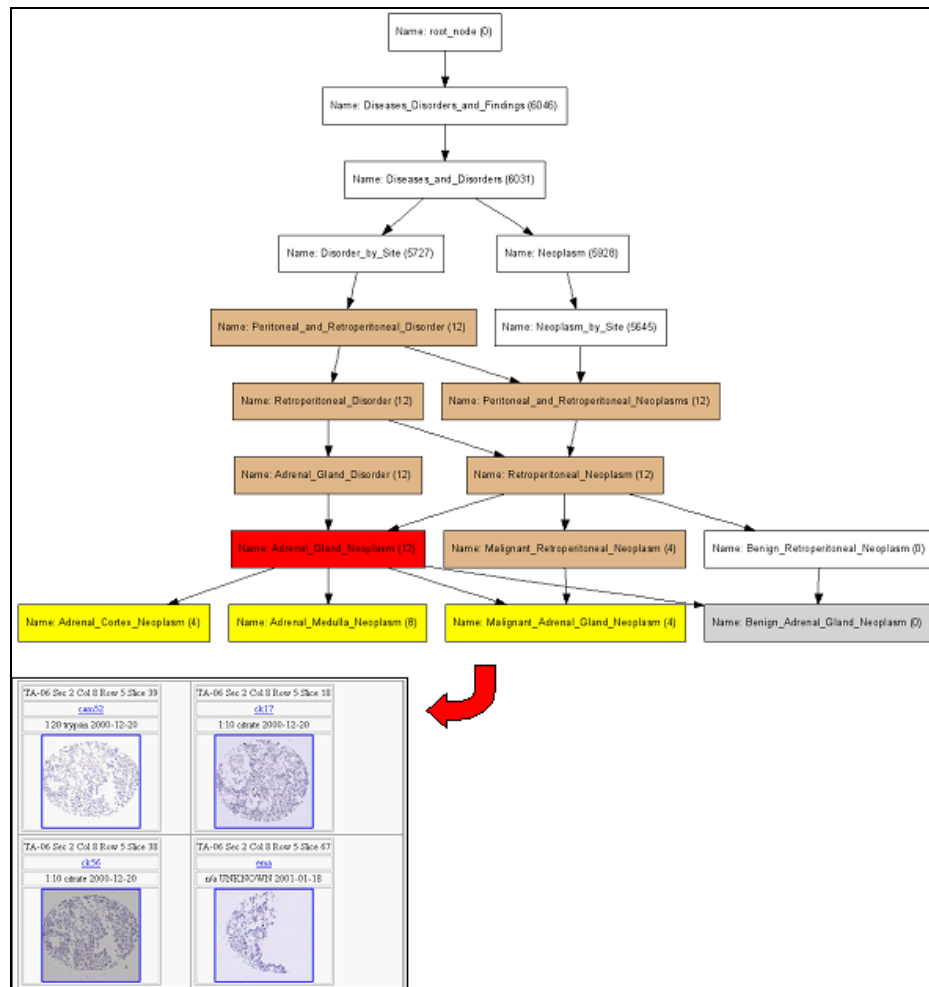


Figure 3 Querying at different levels using an ontology: The figure shows a zoomed in region of the directed acyclic graph view resulting from searching for the term Adrenal gland neoplasm. The red node is the term that has been searched for and then clicked by the user, the yellow nodes are the child terms that have at least one sample in the database assigned to that term, grey nodes are child terms with no corresponding samples in the database and burlywood nodes are parent terms with less than 50 samples. Samples can be retrieved for the selected node.

**3. Encoding Clinical Data in Electronic Medical Records:** A common use of ontologies in clinical medicine is to provide a common way to describe patient information in health records. A comprehensive source for clinical terminology is the Unified Medical Language System, a large collection of biomedical vocabularies developed by the US National Library of Medicine [18]. The UMLS integrates many different biomedical vocabularies, such as the NCBI taxonomy, the Systematized Nomenclature of Medicine (SNOMED), and the International Classification of Disease (ICD). Many of the source terminologies in UMLS are being used in health information systems to provide terms for patient findings, diagnoses, laboratory results, and pathologies [19].

**4. Standardizing the Language for Describing Biomedical Images:** Biomedical images are a challenging type of information to exploit in large repositories because their contents are not explicit. Ontologies can be useful for providing names for the anatomy, pathology, and observations in images. By associating

these names with images (or regions within images), it is possible to analyze large image repositories in terms of these explicit image characteristics. Several ontologies are in development to provide terms for naming entities associated with images. RadLex is a controlled terminology for radiology, providing terms for the techniques, findings, and diseases associated with medical images [20]. The Biomedical Informatics Research Network (BIRN) is creating ontologies to provide the necessary names for interrelated concepts contained in images as well as in distributed online databases [21]. The Open Microscopy Environment (OME) is an open source platform for image microscopy that coordinates the organization, storage, and analysis of the large volumes these data, enabled by the standard terms being drawn from UMLS [22]. The use of ontologies for naming entities in the images permits these projects to unify image data and non-image data, as well as streamline search in large image repositories.

## 2) REPRESENTATION OF ENCYCLOPEDIC KNOWLEDGE

Using bio-ontologies as a source for standardized names is perhaps the simplest use of this technology, but it does not utilize the expressive power of ontologies for representing knowledge through rich relationships. Many applications need to access the knowledge-rich content of biomedicine. Many textbooks have been written to describe the components making up living systems (the entities) and how they work and interact with other components (the relations). Describing complex knowledge in texts makes that knowledge accessible to humans, but not to machines. Ontologies are increasingly being used to structure and make explicit encyclopedic biomedical knowledge in a form that is accessible to both researchers and machines.

### GENERAL DESCRIPTION: FOUNDATIONAL MODEL OF ANATOMY

The Digital Anatomist Foundational Model of Anatomy (FMA) [23] is a comprehensive ontology of human anatomy. The FMA contains more than 70,000 entities that describe the elements of canonical human morphology, providing declarative descriptions of detailed anatomic structures. It is a “reference ontology” in that it specifies canonical knowledge for the domain of anatomy, in the form of a comprehensive set of entities and a large set of rich relationships (*Figure 4*).

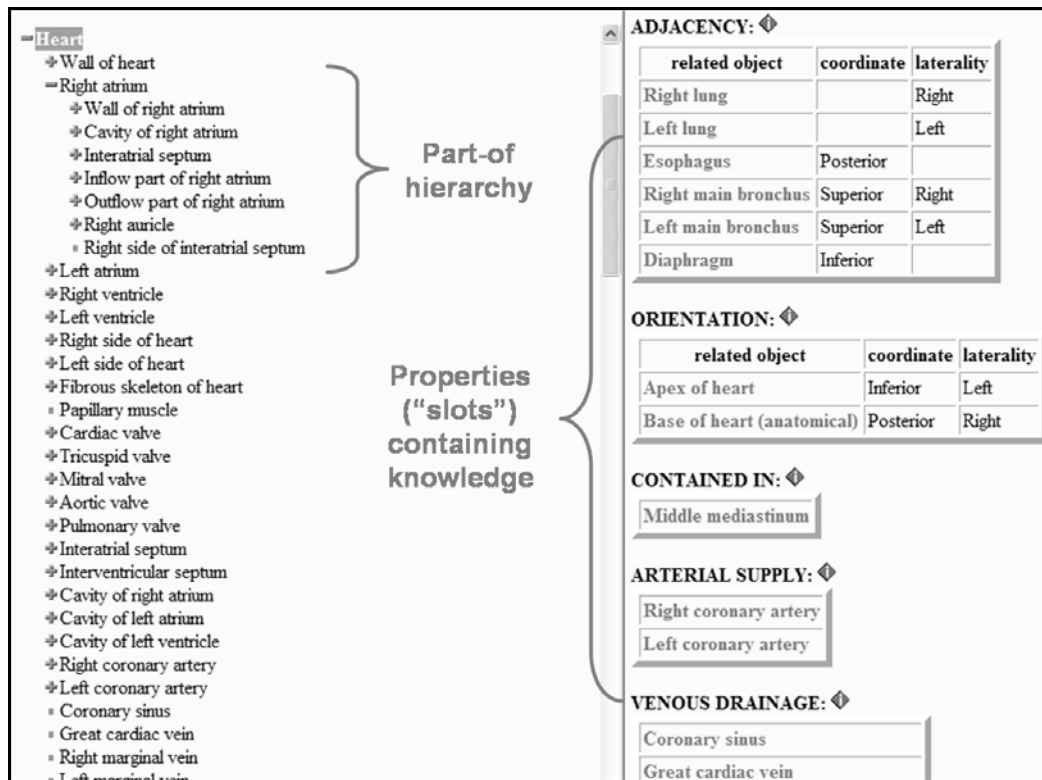


Figure 4- Foundational Model of Anatomy: The FMA is an ontology representing detailed anatomic knowledge. A screen shot of the FMA (accessible by the FM Explorer on the Web at <http://fme.biostr.washington.edu:8089/FME/index.html>) shows that anatomic knowledge is modeled by specifying a large set of rich relations among the anatomic entities. For example, it can be seen that the heart (left) has many relationships to other entities in the FMA (right), such as adjacency, orientation, containment, and vascular supply. Specifically, FMA tells us that the heart is contained in the middle mediastinum, and that it is supplied by left and right coronary arteries.

The FMA was created through disciplined modeling of the structural organization of the human body in collaboration with anatomists and knowledge engineers. It was not created with a particular application in mind; rather, the goal was to provide an electronically-accessible encyclopedic reference for anatomic knowledge.

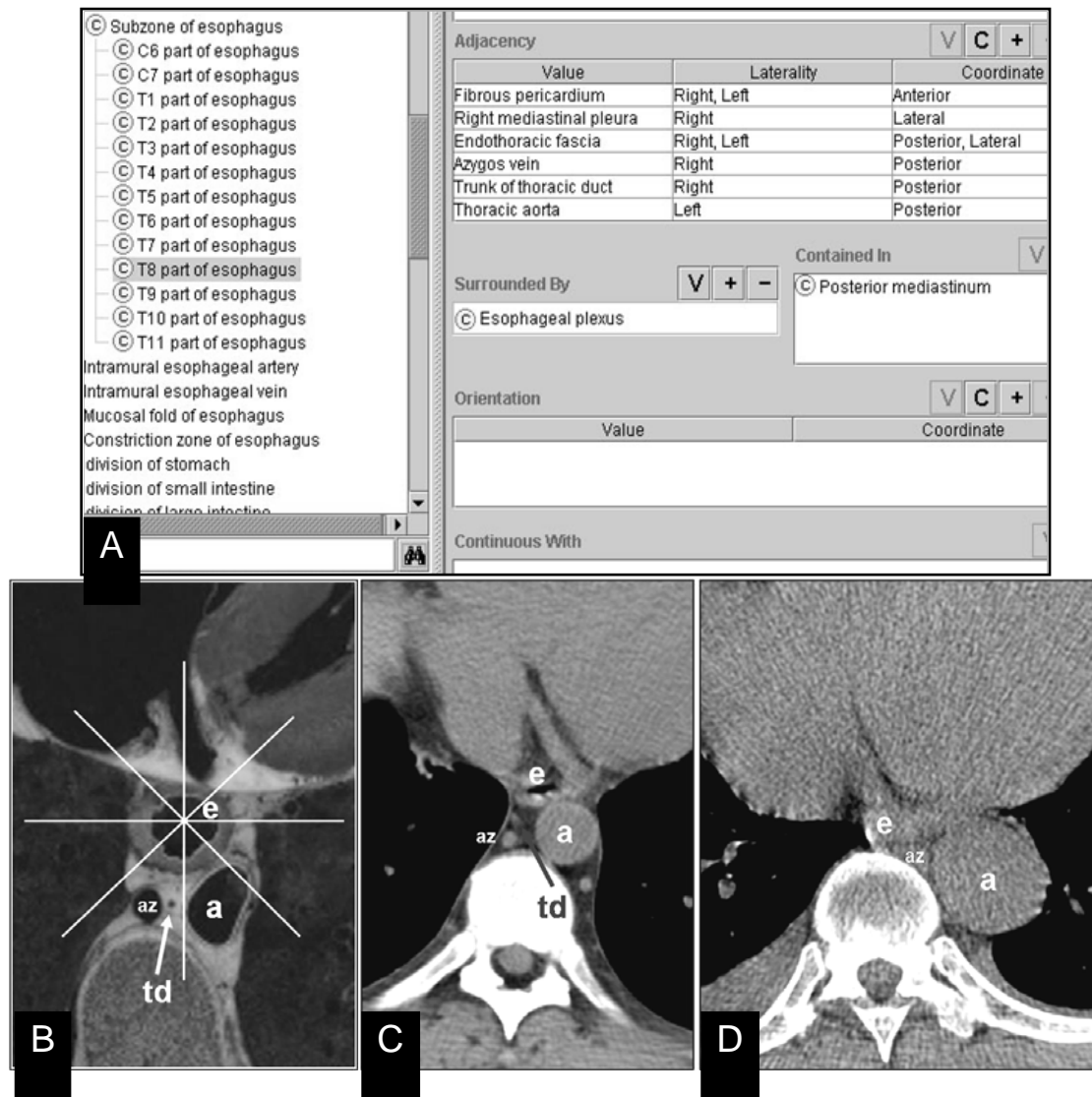
#### APPLICATION

The knowledge in FMA can be exploited in applications that require more detailed information about entities beyond their name. Software applications that need anatomical knowledge about particular organs can access FMA as a “reference ontology,” looking up entities and their relations in FMA to determine canonical facts about anatomic structures needed by an application, such as organ composition, continuity, and adjacency (Figure 4).

For example, an application could be created to use anatomic reference knowledge in the FMA to help radiologists interpret imaging studies, informing them about the anatomic structures affected by abnormalities in adjacent organs (Figure 5). Radiological imaging interpretation is a problem area where anatomic knowledge is needed, but access to that information is limited, because anatomy is incompletely visualized in imaging procedures. Some anatomic structures are not visible in radiology



images due to limited spatial resolution or to individual patient characteristics (Compare *Figure 5C* and *Figure 5B*). A software application can use FMA to recognize that small anatomic structures such as the thoracic duct are adjacent to larger, visible structures such as the esophagus (*Figure 5A*), and inform the radiologist that an abnormality such as a mass in the esophagus may be affecting the adjacent thoracic duct, even though the latter is not visible in the radiographic image (*Figure 5D*). Such detailed anatomic knowledge is useful in informing practitioners about diagnostic possibilities that might have been overlooked.



*Figure 5- Using ontologies to enable knowledge-based applications: Among the rich relations contained in the FMA is knowledge about anatomic adjacency—specifications about which anatomic structures are adjacent. Knowledge about adjacency can be used by a computer application to determine which anatomic structures in the vicinity of abnormality may be affected. (A) A portion of FMA in Protégé showing that detailed adjacency information is represented; in particular, it can be seen that at the T8 level of the esophagus (highlighted class in left panel), the thoracic duct is located to the right and posterior to the esophagus (value for “adjacency slot” shown in right panel). (B) an image from Visible Human at the T8 level of esophagus showing how adjacency in FMA is established using a relative*

*coordinate system (td=thoracic duct, az=azygous vein, e=esophagus, a=aorta; this cross section is viewed from below, such that the left side of the patient is on right side of the image). (C) An axial Computed Tomography (CT) scan at the same level as (B) showing similar adjacency relations as represented in (A). (D) In this CT scan in a different patient from (C), the thoracic duct is not visible, but its presence and location can be deduced from the FMA (A), and this knowledge used to infer that it may be affected by an abnormality (such as a mass) in the adjacent esophagus.*

It has been previously shown that FMA can be useful as a reference knowledge source to predict the anatomic consequences of penetrating injury [24]. In that work, an application was developed to deduce all the anatomic structures that could be injured consequent to penetrating trauma—whether those structures were directly in the path of injury or very close to it. In order for these inferences to be made, the software application queried FMA to find the classes associated with organs that were directly on the path of injury as well as those adjacent to it, informing care takers about organs that are injured as well as potentially injured. Such inferences are made directly from the canonical anatomic knowledge contained in the FMA reference ontology. Another application of FMA's rich knowledge about the anatomic structure of biological systems is as a reference ontology for organizing other biomedical information, including normal and abnormal functions. [25, 26].

#### GENERALIZATIONS TO OTHER ONTOLOGIES AND APPLICATIONS

The use of ontologies as a reference for applications to obtain knowledge has general applicability to many biomedical domains, ranging from helping basic research, to assisting with medical decision support and teaching. For example, Dameron and colleagues developed a numeric and symbolic representation of brain cortex anatomy as a reference ontology [27]. Their ontology and knowledge base could be reused in various application contexts such as teaching, decision support in neurosurgery, and sharing of neuroimaging data for research purposes. Similar to FMA as described above, their reference ontology can drive a broad range of applications that would require detailed knowledge about brain morphological features—that knowledge being obtained by querying the various relations connecting classes in the ontology.

In recent work Kahn and colleagues, created a reference ontology for radiology imaging procedures [28]. This reference ontology contains detailed knowledge about how radiology imaging procedures are performed and the types of images that are produced. Their work demonstrated that diverse intelligent applications could be created using the same ontology as a knowledge source.

### 3) SPECIFICATION AN INFORMATION MODEL (DATABASE/KNOWLEDGEBASE SCHEMA)

Information models outside of the biomedical domain are rarely specified using ontologies; often UML models, entity-relation diagrams, or database schema diagrams are used. However, the use of ontology for the building of models of biomedical information and databases offers several advantages. First, ontologies provide an explicit specification of the terms used to express information in the biomedical domain. Secondly, ontologies enable additional capabilities, such as making relationships among data types in databases explicit, and supporting automated reasoning, such as deducing subsumption among classes. In addition, complex database and knowledgebase schemas are often viewed in a more intuitive manner in ontologies, especially since some ontology tools such as Protégé (<http://protege.stanford.edu>) contain visualization tools that enable developers to create graphical visualizations of schemas [29]. A particular benefit of using ontologies for creating information models is the ability to reuse existing ontologies by inclusion in creating those models. Finally, representations of information models using

ontologies can be published on the Semantic Web if the ontology is in the format of the Web Ontology Language (OWL) [30].

#### GENERAL DESCRIPTION: MAGE-OM, MAGE-ML, MGED ONTOLOGY

Microarrays are currently a very popular experimental method being used to generate molecular-level biomarkers for a variety of biological states and medical diseases. The creation of large amounts of microarray data and the creation of databases to enable sharing of these data quickly raised the need for standards in describing microarray experiments and results. The MIAME standard specifies the minimum information needed to describe a microarray experiment, and the Microarray Gene Expression Object Model (MAGE-OM) and markup language MAGE-ML provide a mechanism for standardized representation of microarray data for data exchange [31]. The MGED Ontology is being created to provide a common terminology and an information schema for annotating microarray experimental results (*Figure 6*). The MGED Ontology provides terms for annotating all aspects of a microarray experiment -- including the design of the experiment and array layout, the preparation of the biological sample and the methods used to analyze the data -- as well as provides a structure for relating these aspects with one another.

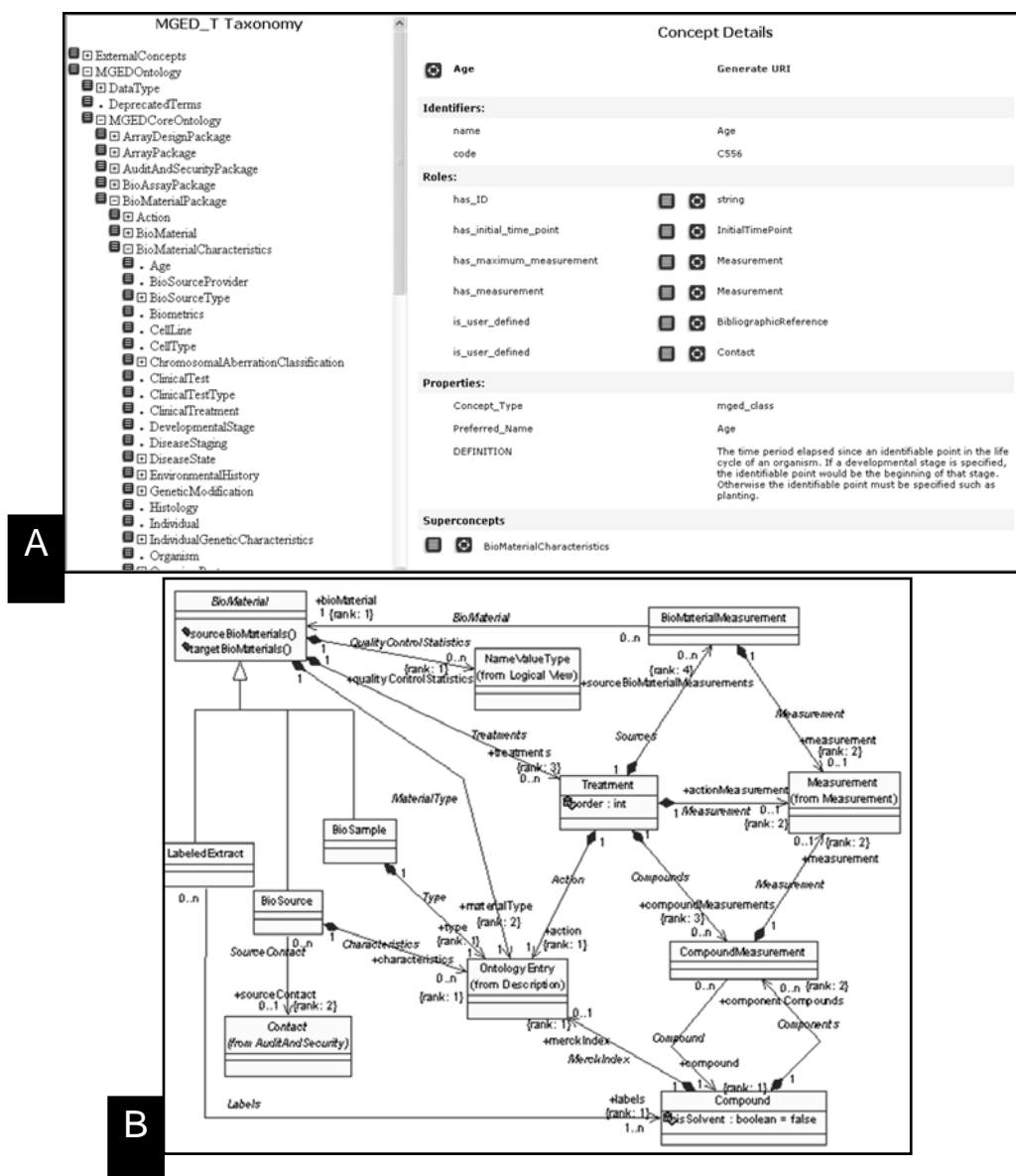


Figure 6- Using ontologies for specifying an information model: (A) This figure shows the MGED Ontology (<http://mged.sourceforge.net/ontologies/MGEDontology.php>). This ontology contains terms that are relevant for describing the design of an experiment, and are used in specifying information models for storing and exchanging microarray data. (B) This figure demonstrates an example information model (MAGE-OM) specified using terms from the MGED Ontology, the latter providing the semantics of entities in the information model.

MAGE-ML is an XML-based markup language that is derived from the MAGE object model, MAGE-OM. MAGE-ML is used to describe and communicate information about microarray based experiments among researchers and microarray databases. MAGE-ML describes microarray designs, microarray manufacturing information, microarray experiment setup and execution information, gene expression data and data analysis results. MAGE-ML is developed and described using the Unified Modeling Language (UML). MAGE-OM is the primary information model for describing microarray experiments.

## APPLICATION

In case of microarray data, not having knowledge of the context in which the experiment is done severely limits the ability to interpret the numbers in the data file. The MAGE-OM, MAGE-ML, and MGED Ontology address the need to specify experimental context by providing a representation of the information models that convey this information. The MAGE-OM describes the information model, and the MAGE-ML conveys the instance data (*Figure 6*). MGED Ontology is used to provide applications with the terms needed to annotate or query microarray data, especially by biologists who may have little knowledge of the ontology structure.

The key advantages of using ontology for specifying information models in the microarray community are to:

- provide standard terms for annotation of microarray experiments when they are submitted to public repositories
- enable unambiguous descriptions of how microarray experiments are performed
- enable structured queries of elements of the experiments, enabling use of MGED ontology to expand queries using the subsumption relations in the ontology structure.

Community microarray database resources such as Stanford Microarray Database, Array Express, and the Gene Expression Omnibus currently require that the submitters of microarray data provide the minimum set of information declared by the MIAME standard.

## GENERALIZATIONS TO OTHER ONTOLOGIES AND APPLICATIONS

The need to represent information models for structured capture of experimental context (under which the experiment was performed), as well as to convey the actual results of experiments to researchers and databases, is not unique to microarray studies. Many other biomedical areas of work could benefit from similar ontology-based structured approaches to describing their data. A large group of researchers have come together to create the Ontology of Biomedical Investigation (OBI), which is creating an ontology for the description of biological and medical experiments and investigations. The ontology will enable the consistent annotation of biomedical investigations. It will model the design of an investigation, the protocols and instrumentation used, the material used, the data generated and the type analysis performed on it; enabling data exchange between researchers and databases similar to the ontology efforts being undertaken in the microarray community.

## 4) SPECIFICATION OF A DATA EXCHANGE FORMAT: BIOPAX

Exchanging data is vital in several biomedical domains where multiple different, yet related databases have emerged over time and they wish to share and link their data. For example, several pathway databases have arisen that contain rich information about metabolic, signal transduction and gene regulatory pathways from particular species. Ontologies, by specifying data exchange formats, can greatly facilitate the process of sharing data amongst biomedical resources such as pathway databases.

## GENERAL DESCRIPTION

Knowledge about biomedical pathways is central to scientific research. There are more than 200 biomedical databases contain information pertinent to pathways. BioPAX is an emerging format for sharing pathways that aims to provide a standard for representing metabolic, biochemical, transcription regulation, protein synthesis, and signal transduction pathways[32]. Pathways in BioPAX are composed of a set of interactions. The top-level BioPAX definition of pathway is general enough to capture the many kinds of pathways used by biologists; figure 7 shows the different kinds of pathway information that can be represented in BioPAX. Pathways in the BioPAX format can be represented as objects using the web ontology language (OWL), which is a language recommended by the World Wide Web consortium (W3C) to create an ontology. OWL can be expressed in RDF/XML syntax or using the N3 triple format.

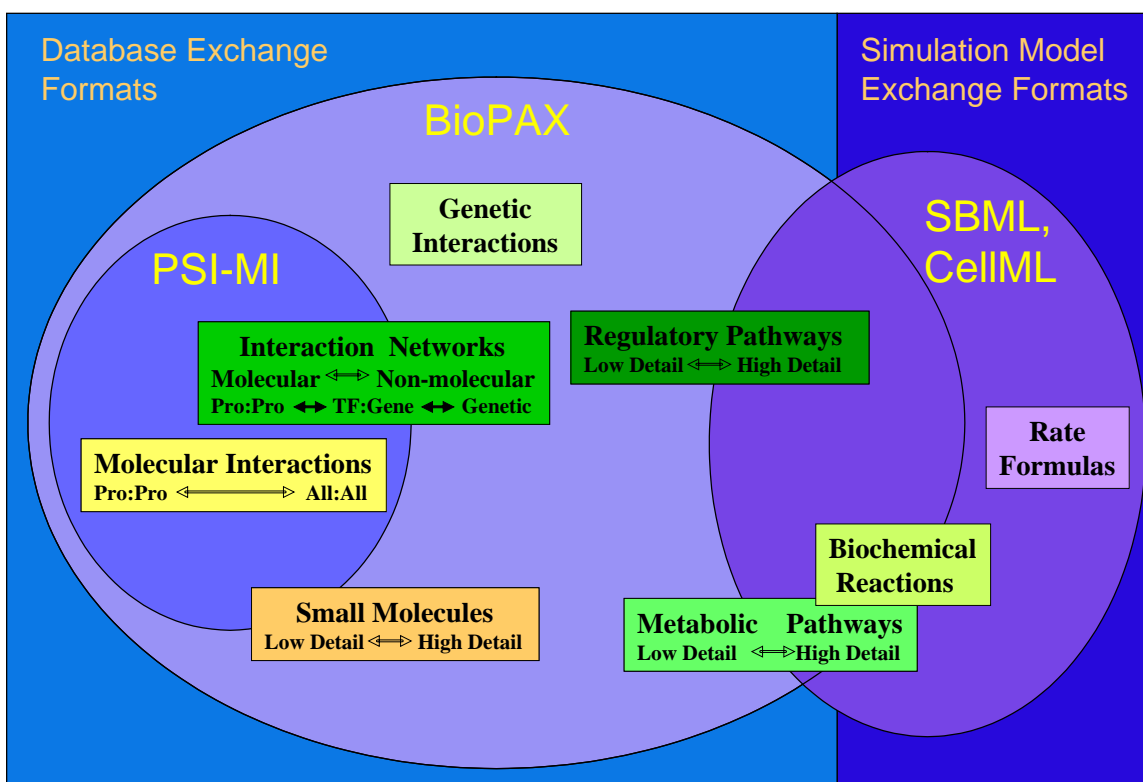


Figure 7- Using ontologies for specifying a data exchange format: The figure shows the scope of BioPAX and the kinds of pathways it aims to represent. At the left are simple binary interactions and at the right are complex models of biological reaction systems that are used for simulation. (Figure courtesy of Gary Bader and Mike Carey)

## APPLICATION

Currently, leading pathway resources such as the Kyoto Encyclopedia of Genes and Genomes (KEGG)[33], BioCyc[34], and Reactome[35] make their data available in BioPAX format and BioPAX viewers are available as additional modules in pathway analysis tools such as PATIKA[36] and Cytoscape[37]. This provides an opportunity to construct unified pathway resources such as the Pathway Knowledge Base project (PKB) that enables querying across different species and across multiple pathway resources simultaneously. It also enables comparison of the degree of complementarity across different pathway sources.

## GENERALIZATIONS TO OTHER ONTOLOGIES AND APPLICATIONS

From the ontological perspective, BioPAX is a very “light” ontology, but therein lays its power. It allows diverse information sources to speak in a simple common language enabling interoperability. In future such common exchange formats will be needed for increasingly exchanging complex artefacts such as simulation models.

## 5) REPRESENTATION SEMANTICS OF DATA FOR INFORMATION INTEGRATION: TAMBIS

The amount of biomedical data available online is enormous, and biomedical discovery commonly occurs by integrating related, yet diverse data from different sources. Performing such integration is laborious, and it does not scale in the setting of the current information explosion.

Ontologies can streamline the process of integrating and accessing data across diverse resources. As described earlier, ontologies provide a means to make the semantics of a domain explicit by providing rich relations among its entities. Specifying the semantics of data in a variety of databases can enable researchers to integrate heterogeneous data across different databases. While it might be simplest to link objects with the same name (syntactical equivalence) in different databases, this is not necessarily a good approach, because names of biological entities (e. g. genes, proteins, pathways) are not the same across databases.

A more robust approach to integrating data is to link based on shared characteristics of biological entities in databases. Ontologies provide a common declarative foundation for describing the content of biomedical databases. Computer reasoning programs can be applied to ontologies to determine if two objects in different databases refer to the same biomedical entity. Thus, an ontology-based framework can facilitate the exchange, integration and validation of information.

### GENERAL DESCRIPTION

TAMBIS is a project that aimed to provide transparent access to disparate biological databases and analysis tools, enabling users to access and virtually integrate a wide range of biomedical resources [38]. Their system includes an ontology (the TAMBIS ontology<sup>2</sup>), a knowledge base of biological terminology (the biological Concept Model), a model of the underlying data sources (the Source Model) and a user interface. The Concept Model provides the user with the concepts necessary to construct queries, and shields the user from the details of the various database sources. The Source Model provides a description of the underlying sources and mappings between terms used in the sources and terms in the biological Concept Model. The TAMBIS ontology serves as a single access point for multiple biological information sources. Queries are phrased in terms of the ontology, and TAMBIS converts them to requests to appropriate sources.

### APPLICATION

The creators of TAMBIS developed an application that uses the TAMBIS ontology to enable users to formulate a query across a set of diverse biomedical sources, providing a means to virtually integrate

---

<sup>2</sup> <http://www.ontologos.org/%5Contology%5CTAMBIS.htm>

these resources (Figure 8). Source-independent, declarative queries formed from terms in the Concept Model are transformed into a set of source dependent, executable procedures.

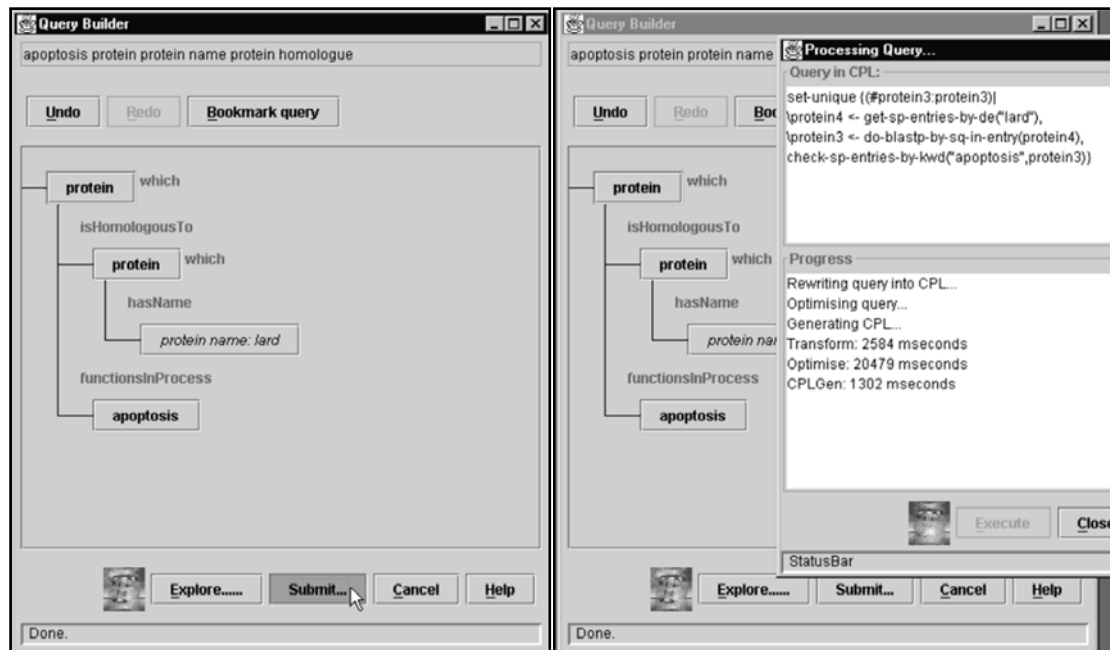


Figure 8- Using ontologies for integrating data resources: Screen shots from the TAMBIS application are shown. A user query is formulated in terms of entities from the TAMBIS ontology (left). In the figure, the query illustrated is: Find proteins which are homologous to the lard protein and functions in the apoptosis biological process.” This query is based on a overarching Concept Model that subsumes all the information models reflected in the databases that the TAMBIS system covers. The Concept Model-based query is translated into appropriate database-specific queries by the application, issued to each appropriate database, and the results collected and returned to the user (right).

The query process in the application proceeds in the following phases:

1. **Query formulation:** the user formulates a query in terms of concepts & relationships in the TAMBIS ontology, constructing a concept describing information of interest using ontology terms. The output of this phase of the application is a source-independent ontology-based conceptual query.
2. **Query transformation:** The application examines the ontology terms comprising the query to identify biomedical database sources needed to answer query, and it then constructs a query plan tailored to the requirements of each source database.
3. **Query execution:** The application submits the individual queries to the relevant source databases and collects the results, returning them to the user. The TAMBIS developers created wrappers for each source database so the latter can be accessed in syntactically consistent manner.

#### GENERALIZATIONS TO OTHER ONTOLOGIES AND APPLICATIONS

The need to integrate diverse data is great in biomedicine, and the methods pioneered in the TAMBIS project are being applied in other domains in biomedicine. Researchers part of the Biomedical Informatics Research Network (BIRN) have recently been undertaking activities to use ontologies to integrate image- and non-image data pertaining to the neurosciences [21]. BIRN employs a “mediator



architecture” to link multiple databases together, each maintaining their specific local schema, into an accessible federated platform. When a researcher queries the BIRN database, the query is processed by the mediator, which uses ontologies to relate and integrate the various source databases. The mediator parses the user query and subsequently submits database-specific queries to the relevant data sources. The approach is thus similar in principle to the approach used by TAMBIS. However, BIRN is currently working on enriching the knowledge contained in the ontology [39], migrating the ontology to OWL, and creating defined classes as well as rich relations that will provide the knowledge necessary for computer reasoning with the integrated data sources (see section on Computer Reasoning below) [40].

## 6) COMPUTER REASONING WITH DATA: HYBROW

Perhaps one of the most compelling advantages ontologies can provide in helping researchers exploit the vast amounts of biomedical knowledge available in electronic form is *computer reasoning*. Computer reasoning encompasses methods that use ontologies to make inferences based on the knowledge they contain as well as any additional contextual information or asserted facts. There is a tremendous amount of biomedical knowledge currently available in online databases and in the published literature. As a result, on the one hand there is an abundance of individual data types such as gene and protein sequences, gene expression data, protein structures, protein interactions and annotations. On the other hand there is a shortage of tools and methods that can handle this deluge of information and allow a scientist to draw meaningful inferences.

Currently, a significant amount of time and energy is spent in merely locating and retrieving information rather than thinking about what that information means. It is extremely difficult to integrate current knowledge about biological systems and formulate hypotheses (or "Models") spanning a large number genes and proteins [41]. It is difficult to determine whether the hypotheses are consistent internally or with data, to refine inconsistent hypotheses and to understand the implications of complicated hypotheses [42]. It is obvious that this situation needs to be rectified and tools need to be developed that utilize formal methods to query and interpret the information at hand [43]. As suggested by Fedoroff et al, we need *tools for thought*: formal representational systems appropriate for representing models of biological systems, and computational tools that can manipulate, check, and use these models to make predictions and form explanations.

### GENERAL DESCRIPTION

HyBrow (Hypothesis Browser) is a system for the representation, manipulation and integration of diverse biological data - such as gene expression, protein interactions & annotations - with prior biological knowledge for the purpose of evaluating alternative hypotheses. Hybrow's purpose is to evaluate and rank hypotheses based on user-defined 'rules', and consistency with all information available to it.[44]

### APPLICATION

HyBrow consists of an event-based ontology for representing hypotheses about biological processes at different levels of detail, a knowledgebase that stores diverse biological information sources such as gene expression, protein interactions & annotations, and programs to perform hypothesis design and evaluation.

- **Ontology for representing hypotheses:** The ontology used in HyBrow allows the representation of knowledge about the Galactose Metabolism in yeast (GAL system) in a manner compatible with the event-based conceptual framework used for reasoning within HyBrow[45]. An event consists of an acting agent (the "subject," such as gene, RNA, protein), a target agent (the "object," such

as a gene, protein, complex), a relationship (the “verb,” such as induce, repress, bind), a context in which the event takes place, and an optional set of associated conditions (such as the presence or absence of other agents) which accompany the event. The construction of events from elements of the ontology, event sets from events, and hypotheses from event sets is governed by a context-free grammar. Events that occur in the same context are combined to form *event sets* and an *hypothesis* consists of event sets linked by logical and temporal operators. An hypothesis must contain at least one event set, which must contain at least one event [44]. Contexts specify *where* events occur in the cell and *under what genetic conditions* they occur. The *contexts* are derived from established ontologies. For example, terms for specifying physical locations in the cell come from the cellular component division of the Gene Ontology. The current hypothesis ontology allows representation of events such as: ‘Gal4p binds to the promoter of the gal1 gene in the presence of galactose in wild type *S. cerevisiae*’.

- HyBrow’s knowledge base stores the different finds of information as well as existing knowledge about the GAL system. The knowledgebase accommodates available literature, curated primarily from YPD[46] at a coarse level of resolution. The knowledgebase was populated by manual curation using loading forms like the EcoCyc database[25] as well as PERL scripting to access the existing public repositories (such as the Saccharomyces Genome Database) to retrieve desired information.
- Hypothesis design and evaluation: There are two interfaces for constructing hypotheses: The diagrammatic interface allows users to draw diagrams using a visual notation constructed in accordance with proposed conventions[27], which are then automatically translated into hypotheses. The widget interface allows the user to construct hypotheses using subject/verb/object selection menus. Hypotheses are saved to local files and then submitted for evaluation via the web. When HyBrow receives an hypothesis, it checks the connections between events and event sets for conformity with the hypothesis grammar. If the hypothesis passes these tests for syntax, each event is then checked for validity using the appropriate rule corresponding to the relationship proposed in the event. For each event, a support, conflict or cannot comment result is returned. Finally, the support and conflict calls are tallied based upon the logical structure of the hypothesis and presented to the user in a web interface [Figure 9]



anatomy that exploits the explicit representation of knowledge in ontologies to support the reasoning process [47].

The FMA ontology of anatomy mentioned earlier has been used in a reasoning application to deduce the physiological consequences of injury to the arteries supplying the heart [48]. In this work, knowledge about which arteries supply different regions of the heart was represented using an OWL ontology, in which the parts of the heart were defined classes, based on specification of the arteries supply each heart region. Connectivity among different arterial segments was specified via continuity relations, and the concepts of arterial occlusion and heart ischemia were defined. A computer reasoning service was implemented that posed the problem of inferring heart injury as a classification problem, based on asserting arterial injuries, classifying the resulting ontology, and reading off newly-classified anatomical entities that reflected the inferred heart injuries [48].

Another way computer reasoning with ontologies has been generalized to other domains is encoding classification criteria in explicit ontology-based form. Criteria for classification abound in biomedicine, and they are generally applied by hand to case data. Ontologies can be used to represent the classification criteria explicit using logic formalisms that some ontology languages provide, such as OWL. In recent work, for example, the classification criteria for interpreting mammography images was represented using OWL to enable creation of applications that automatically classify the interpretation of these studies [49].

In theory, computer reasoning could be applied to knowledge bases created from natural language processing (NLP) methods applied to the biomedical literature to help researchers make sense of it. At the current time, most work has focused on using ontologies to guide NLP according to pre-determined knowledge models or to infer ontologies from text. The GeneWays project developed a knowledge model that enables analysis of signal-transduction pathways in eukaryotes [50]. This system uses the ontology as a knowledge model to analyze the interactions between molecular substances, integrating information extracted from multiple sources by NLP techniques to infer a consensus view of molecular networks. The application provides automatic retrieval of signal transduction data from electronic versions of scientific publications using natural language processing techniques. It also provides visualization and editing of representations of regulatory systems. As the accuracy of information extraction in such systems improves, we will likely see computer inference applications with the biomedical literature.

## DISCUSSION

In this part of the tutorial, we have argued that it is useful to survey biomedical ontologies from a functional perspective—in terms of how they are used. We believe that this approach is helpful to those coming into the field to get a sense for the spectrum of potential use cases and to recognize opportunities for ontology to help with their particular use case. However, researchers and potential ontology developers may face challenges in getting started using ontologies in their work.

## FINDING ONTOLOGIES

The first challenge is that one needs to find existing ontologies that may be suitable for a project, or to select among available ontologies. Over time, the number of ontologies has expanded tremendously. While a few years ago, researchers in particular biomedical domains needed to keep track of one or two ontologies, currently the number of ontologies pertinent to researchers has expanded greatly. Even if an ontology is in a different domain, it could contain much content that is similar to other domains; for

example, mouse anatomy is relevant to workers in the human anatomy domain because there are similarities among many anatomic structures. The proliferation in biomedical ontologies with little concomitant infrastructure development to enable the community to access them has fragmented the field, and the biomedical community is finding it difficult to effectively access and use these valuable knowledge resources.

The National Center for Biomedical Ontology has recently created BioPortal, a Web portal to biomedical ontologies that provides users and software agents comprehensive access to a virtual library of ontologies [51]. The BioPortal ontology library contains over 50 ontologies, including those from the biological and medical domains. The BioPortal Ontology Library unifies ontologies and provides a common access mechanism to ontologies regardless of their original format.

The BioPortal enables users to browse the ontology library to quickly find groups of ontologies related to a variety of domains of interest as well as enables searching of the BioPortal ontology library. Once particular ontologies of interest are located, individual ontologies can be viewed as an expandable tree or graph view (Figure 10).

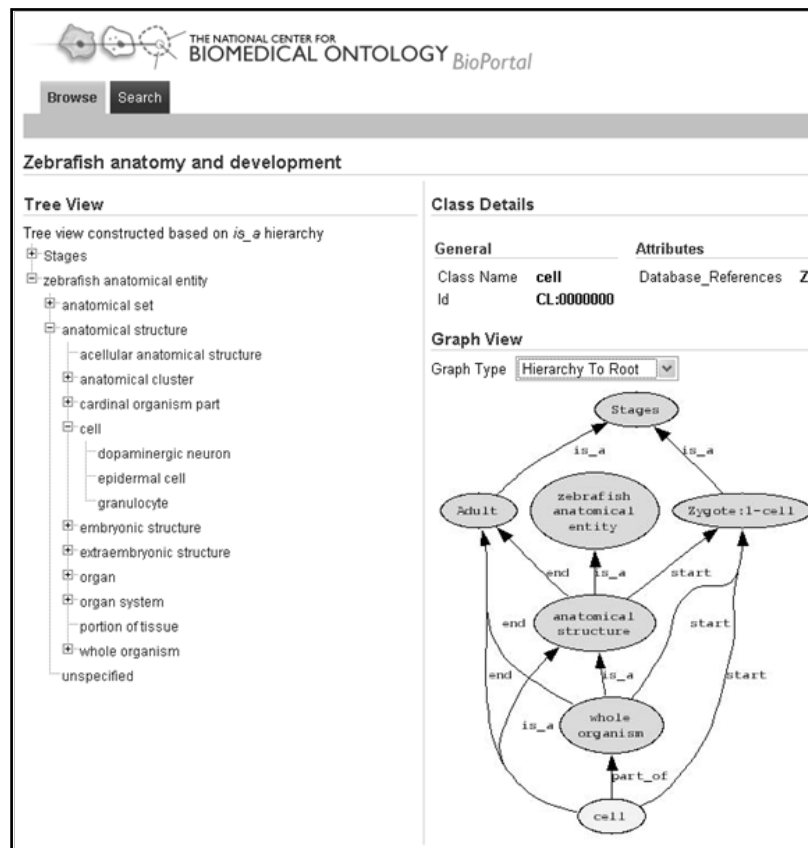


Figure 10- BioPortal: Ontology access and visualization: In BioPortal, ontologies are shown both as an expandable tree (left) as well as a local neighborhood graph (right; selected class is highlighted in yellow).

*The BioPortal ontology library services are provided to external clients as a suite of Web services, enabling developers to access functionality such as ontology categorization and description, graphical ontology browsing, and term search in their applications. This Web portal approach appears promising for unifying and disseminating ontology content, while reducing community fragmentation and providing them the tools needed to exploit these rich resources.*

## KEY POINTS

1. There is a growing community interested in using and producing biomedical ontologies.
2. The diversity of ontologies and their content can bewilder those not already deeply familiar with the field; it is helpful to organize ones thinking about bio-ontologies from a functional perspective.
3. Ontologies are used in biomedicine for naming entities, providing a reference to encyclopedic knowledge, specifying information models, data exchange formats, and semantics of data for information integration, and enabling computer reasoning with biomedical data.
4. The growth in biomedical ontologies has created new paradigms for people to work with them, as well as created the need for new tools to enable collaborative ontology development.

## SECTION II: ONTOLOGIES – WHAT THEY ARE AND WHAT THEY ARE NOT

**Goal:** *In this section, we will discuss “ontology” as understood in philosophy, computer science and information science to explain how the computer/information science meanings are different - but related to - the philosophical meaning of ontology. In practice, ontologies provide standardized labels which are used to annotate different experimental data. We will discuss the implications of this annotation-based view of ontology to clarify the difference between terminologies, taxonomies, application ontologies and reference ontologies.*

### WHAT IS AN ONTOLOGY?

Ontology means different things to different people and we spend a considerable amount of time reconciling them. Here are the most common claims to the "meaning" of ontology.

**Philosophy:** Ontology is the study of what entities and what types of entities exist in reality. [Alt - An ontology is a declaration of the entities & relationships that can exist in a portion of reality (which is of interest to us)]

**AI:** An ontology is a explicit specification of concepts & relationships that can exist in a domain of discourse

**Formal Methods:** An ontology is the statement of a logical theory

**CS:** an ontology is a data model that represents a domain and is used to reason about the objects in that domain and the relations between them

The common ground is that an ontology is a specification of entities (or concepts), relations, instances and axioms in an area of study. Though it fun to argue about these different points of view, in practice that quite pointless unless the question of: **What is an Ontology for and what are you going to do with it**, is answered. Several artifacts are collectively referred to as “ontologies”, the figure below illustrates that there is a continuous spectrum from glossaries at one end and general logic at the other.

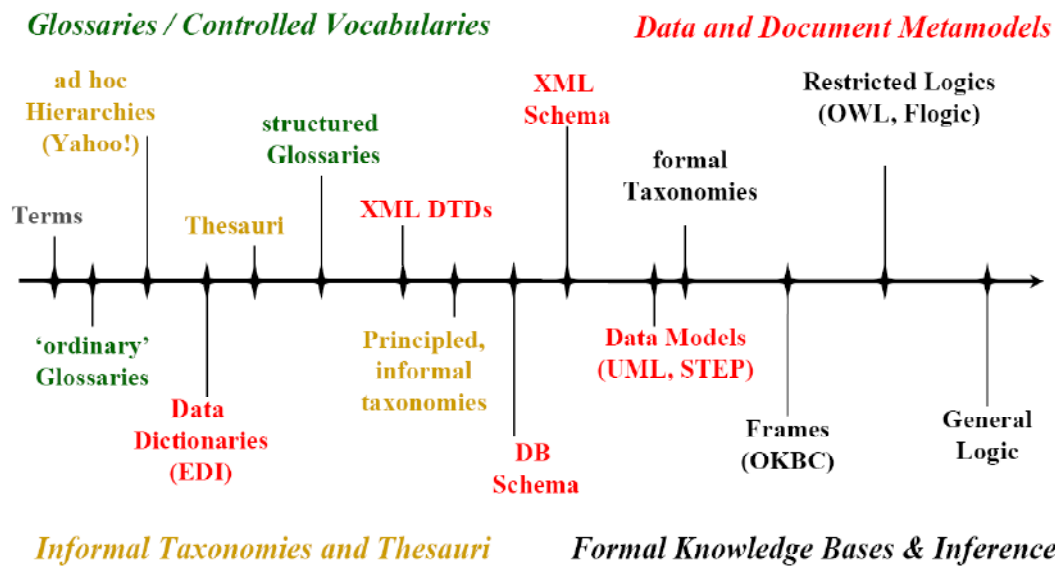


Figure 11: the spectrum of knowledge artefacts. (Originally by Michael Uschold, used with permission)

## APPLICATION ONTOLOGY VS. REFERENCE ONTOLOGY

A reference ontology is analogous to a scientific theory, it consists of representations of biological reality which are correct according to our current understanding.

An application ontology is a software artifact for structuring data according to some hierarchy of classes, for the purpose of managing and manipulating that data, supporting interoperability of various resources.

We believe that as far as possible, one should focus on developing scientific information models, data-models, and process-models etc that are as close as possible to and that refer to *reference ontologies*.

## WHAT AN ONTOLOGY IS NOT

- An ontology is not the same as a knowledgebase:
  - Ontology (types) + Instances = KB
- An ontology is not the same as a database schema
  - A database schema is designed to store the instances conforming to an ontology
- An ontology is not the same as an XSD
  - An XSD tells you how to structure the information that describes the instances



## TRADEOFF BETWEEN SEMANTIC RIGOR AND COMPLEXITY OF REASONING

While it is tempting to be as rigorous as possible while creating an ontology, we would like to stress again that the key question is: **What is an Ontology for and what are you going to do with it.** Developing knowledge structures with rigorous semantics is hard, time consuming and expensive. The need to do that has to be weighed against the complexity of reasoning that is required for the pertinent use case. For example, if the end goal is database cross-linking then it is probably a waste of time to build in strong semantics that support modal logics. The figure below, illustrates the balance between the complexity of reasoning possible and the semantics needed for it

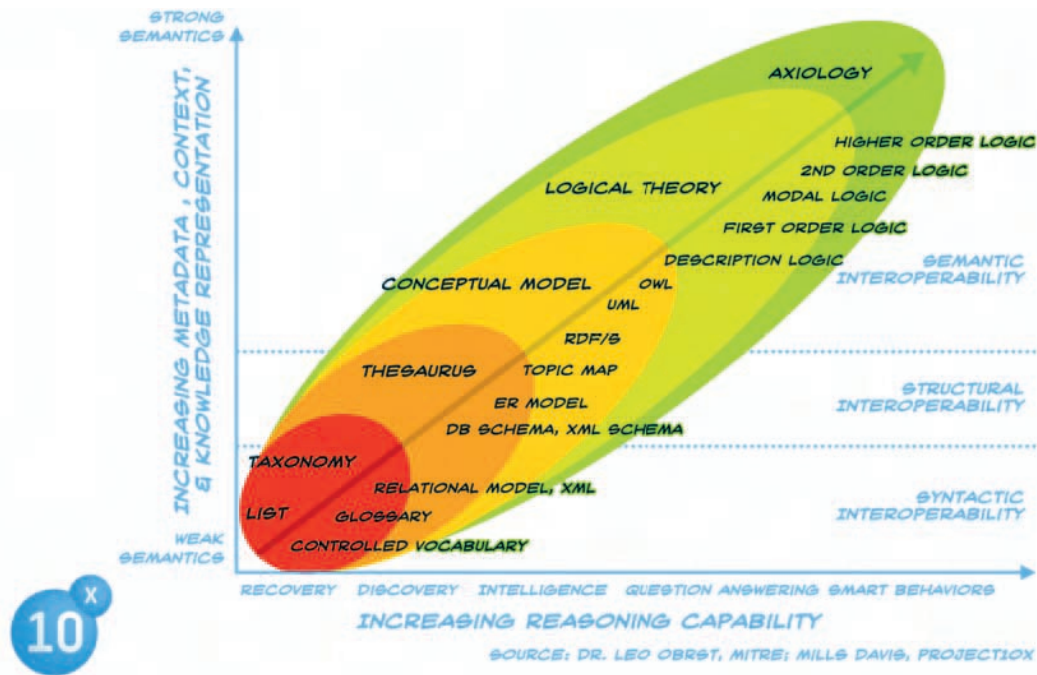


Figure 12: Relationship between semantic rigor and complexity of reasoning. The red zone indicates the most 'bang for the buck' zone where there is maximum gain with the least effort. Developing strong semantics that support first order logic (or more expressive logics) has to be justified by demonstrating a strong need for that.

## SECTION III: BUILDING ONTOLOGIES USING OWL

**Goal:** *We will provide an overview of the various constructs available in OWL and how they can be used to represent biomedical knowledge in an ontology. We will demonstrate an example of how DL-reasoners can be used to verify ontologies created in OWL and how such reasoning will help in reducing errors in biomedical ontologies. We will provide indications of some of the problems facing ontology developers using OWL in the bioinformatics domain.*

Weaving the  
Biomedical semantic web

Participants are requested to read the included published paper, *weaving the Biomedical semantic web with protégé-owl*, for details on this topic

## **SECTION IV: INTERACTIVE SESSION: DEVELOPING A SMALL ONTOLOGY**

**Goal:** *This session will consist of an audience-driven discussion with the goal of developing an ontology to represent DNA sequences. We will focus on the task of representing relationships between introns, exons, start sites, promoters, genes and chromosomes. This domain is accessible to participants with a wide background and has enough complexity that a large fraction of the common mistakes do happen. The mistakes made will be utilized to reinforce the best practices that follow.*

Session Notes

## SECTION V: THE DO'S AND DON'TS OF ONTOLOGY CREATION

**Goal:** *In this section, we will describe the key principles of ontology design that address the most common sources of mistakes in biomedical ontologies. We will then discuss rules of thumb for good ontology design, along with examples of how they will enhance the quality of the resulting ontologies. We will also outline the guiding principles of the OBO Foundry, and describe how they might enable cross-ontology reasoning. (The text in this section is based on work of Andrew Spear and Barry Smith)*

Once the domain information has been assembled in the form of recognizing the important domain entities and relations, and the appropriate scope, granularity and the level of semantic rigor required for the ontology have been determined, the next step is to organize the domain information in a systematic and coherent fashion. **The goal of this organization is to develop a representational artifact that is as logically coherent, unambiguous and true to the facts of reality as possible.**

### ORGANIZING THE ENTITIES

There are three major and interrelated facets of organizing the entities for domain ontologies. These are 1) terminological regimentation, 2) definition of terminology, and 3) construction of taxonomic hierarchies of terms based on *is\_a* relations.

#### TERMINOLOGY

The process of gathering domain-specific information results in the construction of a lexicon or terminology. However, if the goal is to use this domain specific information for purposes of representation in a computer-based ontology, then a more rigorous formalization – regimentation – of this terminology is needed. Regimenting a terminology thus involves both the explicit statement of (as well as ruthlessly consistent adherence to) syntactic conventions for the writing of terms and explicit consideration of the intended audience or user-base for the ontology. Specific principles along these lines include the following.

#### TERMINOLOGICAL CONSERVATISM

Don't reinvent the 'wheel': There are already a sufficient number of words in the world and in the biological and medical communities to ensure that the creation of new highly **specialized terms for purposes of inclusion in a domain ontology will rarely, if ever, be necessary**. The terminological choices of domain ontology builder(s) should be as respectful as possible of the current terminology, usage and practice of contemporary domain experts and potential users of the ontology.

A simple principle to follow, in selecting terms for a domain ontology, is to stay as close as possible to the actual use of people working in the field the domain ontology is about. Terms that are widely used and well-known by domain experts should be given preference over highly specialized and little used terms, and given the same meaning that they currently have in their use by domain experts. Creating new terms to represent things that a community is already familiar with or using a familiar term with a new and different meaning, are both likely to lead to confusion – both in the encoding of information into the ontology, and in its interpretation by end-users.

#### SINGULAR NOUNS

For the sake of intelligibility: **the general terms in an ontology should be formulated in the singular**, and the ontology's documentation should pay careful attention to the distinction between singular and plural

nouns and to the requirement of noun-verb agreement. Thus ‘cat’, not ‘cats’, and ‘eukaryotic cell’, not ‘eukaryotic cells’.

There are a number of reasons why this convention should be adopted. First, it is crucial that some syntactical standard or other be adopted and rigorously adhered to for the encoding of common nouns, in order to ensure that they always appear similar to human users. In this respect rendering all such terms in the singular is as good a decision as any. Additionally, ensuring grammatical intersubstitutability of terms with their corresponding definitions (something that will be further discussed below) will be much easier if all terms have a standard grammatical format. Second, a more principled reason for representing the common nouns or universal terms in an ontology in the singular, is that the common nouns in an ontology always refer either to universals or to defined classes. In either case, however, the reference of these terms is singular. There is only one universal “feline”, even if it has many instances, and there is only one defined class “all the debutants in Texas in 1984”, even if it has many members. Thus it makes sense to use singular rather than plural terms to refer to entities such as universals and classes, and to do this consistently when constructing a domain ontology.<sup>3</sup>

#### COMMON NOUNS IN LOWER CASE:

For the sake of intelligibility: **represent terms referring to universals or classes in all lowercase letters.** Thus ‘cat’, not ‘Cat’ or ‘CAT’, and ‘eukaryotic cell’, not ‘Eukaryotic Cell’ or ‘EUKARYOTIC CELL’. As with the convention regarding use of singular nouns, this convention is proposed largely because some convention or other must be adopted and rigorously consistently adhered to. However, in English capital letters are normally used to indicate either a proper name (Tom, Seattle, Jupiter) or an acronym (the U.N., the E.U., the U.K.), whereas common nouns normally do not involve capital letters of any sort. It is thus more consistent with English usage to use all lower case letters for the encoding of general terms.<sup>4</sup>

#### AVOID ACRONYMS

For the sake of intelligibility: **don’t use acronyms as part or all of any term.** Thus, instead of ‘ATP’ or ‘atp’ write ‘adenosine triphosphate’, and instead of ‘dna’ or ‘DNA’ write ‘deoxyribonucleic acid’. Using an acronym rather than the term for a universal or class increases the chances of confusion on the part of domain expert-users, while rendering use of the ontology by non-domain experts nearly impossible. In the worst case, an ontology whose terminology is filled with acronyms will be equivalent to an ontology intended to be used by speakers of French that is written in Russian. The ontology itself will not be understandable or usable without some further interpretative or translational guide.

#### UNIVOCITY

For the sake of intelligibility: **Terms should have the same meaning on every occasion of their use.** In an ontology, ‘cell’ should mean cell, ‘cancer’ should mean cancer, and similarly in all other cases. The principle of univocity in ontology terminology development is difficult to maintain because ordinary

---

<sup>3</sup> Barry Smith. “[Against Idiosyncrasy in Ontology Development.](#)” Forthcoming in B. Bennett and C. Fellbaum (Eds.), *Formal Ontology and Information Systems*, (FOIS 2006), Baltimore November 9–11, 2006.

<sup>4</sup> There are of course other languages with other grammatical rules for capitalization of nouns. Should it prove more intelligible to use capital letters in an ontology the natural language of which is, for example, German, then by all means this convention can be altered. Again, what is crucial is that the convention used by an ontology be explicitly stated and consistently adhered to throughout.

language regularly violates it. For example, the English word 'bank' can mean both "a financial institution" and "a stretch of earth directly connected and running parallel to a river".

The reason for avoiding such ambiguity in the context of ontology design is quite straightforward. If a single term is used in more than one way in a given context, human participants in discourse regarding that context as well as computer applications working under different contexts are likely to become confused; leading to both computational errors and user confusion.<sup>5</sup>

### Universal/Instance Univocity:

For the sake of intelligibility: **Terms/expressions referring to Universals, and terms/expressions referring to instances should be clearly demarcated.** For example, the common noun 'dog' can be plausibly understood as referring to a type or universal "dog". The term 'dog' which occurs in the sentence 'The accident was caused by a dog unintentionally ejected from a motor vehicle due to failure to use restraining harness' can be plausibly understood as referring to a single particular dog. These two uses of the term 'dog' should be kept clearly separate in an ontology. There are a number of different ways to do this. One simple way would be to abide by the conventions we have already put forward for representing common nouns, using 'dog' to refer to the universal dog, while using a capital letter, proper names or alphanumeric strings to refer to particular dogs, as in 'Dog', 'Fido' or '#d437' (importantly, by using one of these conventions consistently, not by mixing all three together!).<sup>6</sup>

### DEFINITION OF TERMS

Regimenting the definitions of terms in an ontology is a semantic task, one that has to do with providing a definitive statement of the nature of the things that the terms refer to. In a scientific ontology we are not interested in the *lexical* or *conventional* definition of a term, a definition that reports the meaning that all or most members of a given language community attribute to a term (as can be found in a dictionary), but rather in the *real* definition of a term, that is, in a scientific statement of the basic nature of the kind of thing that that term refers to. It is important that the definitions be as clear, consistent and accurate as possible, while also being organized in terms of a coherent and consistently applied set of conventions. In the following we put forward a number of principles for the formulation of domain-ontology definitions.

### ESSENTIAL FEATURES

Use essential features in defining terms: **The definition of a term referring to a universal or kind should be stated in terms of the essential features of the entities that are instances of that kind.** The essential features of a thing are those features without which the thing would not be the kind of thing that it is. For example, it is arguably not essential to a thing's being an instance of the universal human that it have precisely two legs, ten fingers, an appendix or blond hair. On the other hand, if an entity is unable to engage in any kind of communication, to think in a (somewhat) rational way, or to have certain kinds of self-reflective or self-aware thoughts, then this might be grounds for maintaining that this thing is indeed

---

<sup>5</sup> For a discussion of the violation of the principle of univocity regarding the relations *part\_of* and *is\_a* in the Gene Ontology, see Smith, B., Köhler, J., & Kumar, A. On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology; see also Barry Smith. "[Against Idiosyncrasy in Ontology Development](#)." Forthcoming in B. Bennett and C. Fellbaum (Eds.), *Formal Ontology and Information Systems*, (FOIS 2006), Baltimore November 9—11, 2006.

<sup>6</sup> Barry Smith. "[Against Idiosyncrasy in Ontology Development](#)." Forthcoming in B. Bennett and C. Fellbaum (Eds.), *Formal Ontology and Information Systems*, (FOIS 2006), Baltimore November 9—11, 2006.

not an instance of the universal human. Thus Aristotle defined 'human' as "an animal that is rational". Taking rationality and being an animal to be essential features of human beings means saying that while many other features of a thing may change (such as hair color, skin color, body parts, height, weight, strength, taste in food, etc.), develop or be eliminated altogether, a thing cannot lack either the feature of being an animal or the feature of being rational and still be an instance of the universal human being.

*For natural objects*, such as those studied by chemistry, biology and physics, the essential features of a thing are usually the features of that thing that play a prominent role in scientific explanation of its existence and behavior. Thus a good Aristotelian definition of 'water' would be water is a molecular compound consisting of hydrogen and oxygen, or water is the molecular compound H<sub>2</sub>O. *For artifacts*, objects created by humans to be used in various contexts, the essential features usually have to do with the purpose or use for which the artifact was created. Thus a knife is a tool for cutting things, while a chair is furniture that can accommodate a normal sitting human being.

**Definitions in a domain ontology should always be in terms of the essential features of the entities under consideration. What is essential to the domain as a whole will, as a general rule, be determined by the statement of the intended scope of the ontology.** A final point is that with regard to defined classes that do not refer to universals on the side of reality, the essential features to be used in the definition just are the features mentioned in the arbitrary designation of the class. Thus the 'essential' features of the class of all people suffering from HIV on the African continent just are "to be suffering from HIV" and to be "on the African continent".

Some examples of definitions that fail to utilize essential features of the things being defined include the definition of 'water' from the International Classification of Nursing Procedures as "a type of Nursing Phenomenon of Physical Environment with the specific characteristics: clear liquid compound of hydrogen and oxygen that is essential for most plant and animal life influencing life and development of human beings",<sup>7</sup> the definition of 'living subject' as "a subtype of Entity representing an organism or complex animal, alive or not" from the HL7 RIM,<sup>8</sup> and the definition of 'person' as "A living subject representing a single human being [sic] who is uniquely identifiable through one or more legal documents", from the HL7 Glossary.<sup>9</sup>

#### ARISTOTELIAN STRUCTURE

Use Aristotelian structure when formulating definitions: Consider again Aristotle's definition of 'human': a human is an animal that is rational. This definition **has the basic form *An A is a B that Cs/is (a) C***; an A (human) is a B (animal) that Cs (is rational). This basic format should be used to structure the definitions that are provided for terms anywhere in a domain ontology. The advantages of using this structure are that A, B, and C will always occupy the same places in the definition, and they can always be interpreted in similar ways, regardless of the specific domain in which terms are being defined.

---

<sup>7</sup> International Classification of Nursing Procedures (ICNP), <http://www.icn.ch/icnp.htm>, accessed May, 2006.

<sup>8</sup> HL7 Version 3.0 accessed via Knowledge Source Server Version 2006AC, Thursday, September 28, 2006.

<sup>9</sup> Various Contributors eds., HL7 Publishing Technical Committee. Last Published 11/22/2005 8:05 PM. HL7® Version 3 Standard, © 2005 Health Level Seven®, Inc.

The traditional Aristotelian definition structure should be understood in the following way: A is the term that is being defined ('human', 'chair', 'cell', etc.), B refers to the genus of the original term, the next highest class/universal in the hierarchy of classes/universals in which the term is located, and C refers to the differentia of the universal designated by A. The differentiae of A are the essential features of A, those features that any entity must possess in order to be an instance of A, and those features that distinguish entities of kind A from all other entities. Thus in the Aristotelian definition of human, a human (A) is an animal (B: genus) that is rational (C: differentia). The structure of Aristotelian definitions can be understood against the background of species-genus hierarchies, the taxonomies that universals naturally form in which the higher levels of the hierarchy represent universals of greater generality (genus) relative to the lower levels (species) in the hierarchy.

One advantage of consistently using the Aristotelian definitional structure is that it can be used to explicitly locate the place in *is\_a* hierarchies of the universals referred to by the terms being defined. The Aristotelian definitional structure thus represents a consistent format for the representation of definitions that can be used regardless of ontological domain, and that is inherently directed at explicitly representing the location of the term defined in an *is\_a* hierarchy based on the informational content and structure of the definition alone.

#### ARISTOTELIAN APPROACH

Define the terms in an ontology from the top down: **Terms in an ontology should be defined by beginning with the most general universals, and then by systematically working 'downwards' towards the least general.** This procedure is highly consistent with the principle requiring the use of Aristotelian structure in definitions. Beginning with an undefined or primitive top node or root term, terms on the next level down will be defined by saying that an A is a B (top level node and genus) that Cs/ is (a) C (differentia). This procedure can be reiterated as many times, and at as many different levels as necessary, but starting from the most general level keeps things simple at the beginning, and gives the ontology developer a better perspective from which to assess the comprehensiveness of the ontology that she is building.

A more general consideration in favor of the top down approach comes from the point, already discussed earlier, that an ontology should have a well defined and delimited domain to which it is intended to apply, one that is determined, as much as possible, by the actual unity of scientific and practical domains of research in reality. Thus beginning with the more general entities and relations in an ontology and working downwards ensures ruling out, from the beginning, consideration of entities that are not relevant to the domain one's ontology is intended to represent.

#### POSITIVITY

Don't use negatives: Definitions in a scientific domain ontology are intended to convey the essential information about their subject-matter to a user. Utilizing negative predicates (non-physical, non-environmental, non-cellular), or negative characterizations (not a part of the heart, not a breathing thing) involves providing much less information about the entities referred to by the term being defined than would be provided if only positive characterizations were given.

Compare, for example, a definition of heart as "an organ that is *not* part of the nervous system", with the definition from the FMA, the heart is an "Organ with cavitated organ parts, which is continuous with the systemic and pulmonary arterial and venous trees". The first, negative, definition of the heart ensures only that the heart is not the brain, while leaving entirely open the possibility that it is the lungs, the



kidneys or any number of other organs that are “not part of the nervous system”. Thus, while negative definitions do provide *some* information about the entities being defined, positive definitions are much more exact and provide much more information, and they should be preferred and formulated whenever possible in the construction of domain ontologies.

#### INTELLIGIBILITY

Keep it simple: **The terms used in a definition should be simpler (more intelligible, more scientifically, logically or ontologically basic) than the term to be defined.** Definitions of terms are given in order to explain to people who do not know the meaning of the term what that meaning is. It is generally the case that a person who does not know the meaning of a term, especially a technical term, also does not know the meaning(s) of terms more abstract or complex than the original term. Thus a definition that uses such abstract or complex terms in defining the original term is unlikely to serve its purpose.

In scientific contexts it is inevitable that definitions will involve a certain degree of complexity and specialized terminology, however this should be kept to an absolute minimum in ontology design. Further, when specialized and potentially obscure terminology is used in the definition of a given term, the ontology should either itself include or at the very least include references to clear definitions of this terminology itself.<sup>10</sup> Some examples of definitions that violate the principle of intelligibility in actual ontologies include the old BIRNLex definition of ‘mouse’ as a “common name for the species *mus musculus*”,<sup>11</sup> and the old ‘GO:0007512: adult heart development’, which was defined as “generation and development of the heart of a fully developed and mature organism”.<sup>12</sup>

#### NON-CIRCULARITY

Avoid circularity in the definition of terms: **A definition is circular if the term to be defined, or a near synonym of that term, occurs in the definition itself.** For example, defining ‘plant cell’ as “a cell that is found in plants” or ‘surgical tool’ as a device that is used in surgical procedures. These definitions are circular because they provide no more information about the nature of the things the terms refer to than the terms themselves provide. Since definitions are intended to explain the meaning of a term to someone who does not already understand that term’s meaning, using the term itself or some very similar expression in its own definition defeats the purpose of providing a definition in the first place.<sup>13</sup>

---

<sup>10</sup> Smith, B., Köhler, J., & Kumar, A. On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology.”

<sup>11</sup> BIRNLex, <http://137.110.143.4:8080/BIRNLex/>. Note: this definition has now been fixed.

<sup>12</sup> Smith, B., Köhler, J., & Kumar, A. On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology. Note: this definition has now been fixed.

<sup>13</sup> See Kohler J, Munn K, Ruegg A, Skusa A, Smith B. “[Quality Control for Terms and Definitions in Ontologies and Taxonomies.](#)” *BMC Bioinformatics*. 2006 Apr 19;7(1):212.

Werner Ceusters and Barry Smith. “A Realism-Based Approach to the Evolution of Biomedical Ontologies.” forthcoming in *Proceedings of AMIA Symposium*. 2006. <http://ontology.buffalo.edu/bfo/Versioning.pdf>, and also Barry Smith. “[Against Idiosyncrasy in Ontology Development.](#)” Forthcoming in B. Bennett and C. Fellbaum (Eds.), *Formal Ontology and Information Systems*, (FOIS 2006), Baltimore November 9—11, 2006.

Some (now repaired) examples include the old GO definition of ‘hemolysis’ as “the causes of hemolysis”,<sup>14</sup> and the old BIRNLex definition of ‘eyeball’ as “the eyeball and its constituent parts”.<sup>15</sup>

#### TERM-DEFINITION INTERSUBSTITUTABILITY

In all extensional contexts **a defined term should be intersubstitutable with its definition in such a way that the result is both grammatically correct and truth-preserving**. The basic idea behind this principle is that wherever a term refers to a thing the definition of that term should also successfully refer to that thing. The intersubstitutability of a term and its definition with regard to the truth-value of sentences in which they occur is important both for preserving truth across inference in automated reasoning contexts and for ensuring intelligibility for human users of ontologies. If replacing a term with its definition results in a grammatically incorrect expression, this will substantially impede the human usability of an ontology.<sup>16</sup>

#### CONTEXT-INDEPENDENCE

Don’t leave the definition of a term open to interpretation: It should not be up to the end user of a domain ontology to decide or interpret whether or not the term ‘heart’ in the ontology means “human heart” or “canine heart”, nor should it be up to the user to decide whether ‘cell’ means “animal cell”, “plant cell” or “cell in general”. This information should be explicitly included, either in the term itself (say ‘plant cell’ rather than just ‘cell’), or in the definition of the term.

Scientific theories are intended to express the truth about reality in their respective domains, full stop. Thus a scientific definition, ideally, is not just sometimes or partially true, but is true, period. Definitions of scientific terms should attempt to capture this fact. Conversely, rendering the definitions of many scientific terms context independent will involve including more information about context in these terms themselves and/or in their definitions. For example, anatomy is the study of the physical structures present in organisms in general, whereas human anatomy, mouse anatomy, etc. are particular sub-fields of anatomy in general. Thus terms and definitions within these sub-fields, in order to be as context free as possible, should include the fact that they are definitions within a sub-field. The following definition of ‘cell’ as “structural and physiological unit of a living organism; it (i.e., plant cell) consists of protoplast and cell wall” from the Plant Ontology, violates this principle because it implicitly characterizes the term ‘cell’, which one would normally expect to refer to the general universal “cell”, as applying only to cells within a specific domain, namely “plant cells”.<sup>17</sup> Here it would be much better if, instead of ‘cell’, the term ‘plant cell’ was used.

#### MODULARITY

A set of definitions should be modular: Modularity is not a feature of a single definition, but rather a property that a set of definitions has if it has been structured in a certain way. A set of definitions

---

<sup>14</sup> Gene Ontology, <http://www.geneontology.org/>. Note: this definition has now been fixed.

<sup>15</sup> BIRNLex, <http://137.110.143.4:8080/BIRNLex/>. Note: this definition has now been fixed.

<sup>16</sup> For an example from GO, see Barry Smith, Jacob Köhler and Anand Kumar. “[On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology](#)” in *Data Integration in the Life Sciences, First International Workshop, DILS 2004*, Leipzig, Germany, March 25-26, 2004, Proceedings, Editor: Erhard Rahm.

<sup>17</sup> Plant Ontology, <http://www.plantontology.org/>. Note: this definition has now been fixed.

satisfies the requirement of modularity if they are organized into levels, with level 0 (root or most general) terms picked out as undefined primitives, and terms on levels  $n+1$  for every  $n$  greater than 0 being defined by referring exclusively to logical and ontological constants together with already defined terms taken from levels less than  $n+1$ . In practice what this rather complicated principle means is just that the terms in an ontology should be defined in Aristotelian fashion, from the most general to the most particular, while making use of the Aristotelian definitional structure, and abiding by the other principles that have already been discussed.

The principle of modularity is thus closely related to the fact that a system of well-defined terms regarding a specific domain should in normal cases form a hierarchically structured taxonomy. More specifically, if all or most of the terms being defined refer to universals on the side of reality, then the hierarchy amongst universals from more specific (cat, fern, human) to more general (mammal, plant, organism) should be reflected in the definitions of the terms that refer to these universals. The principle of modularity is explicitly intended to ensure that terms lower down in a taxonomic hierarchy inherit all properties and characteristics from their parents, and following this principle helps to ensure logical consistency in the definition of terms, clear demarcations amongst levels of abstractness within the ontology, and the possibility of automated reasoning.<sup>18</sup>

#### TERM-DEFINITION/TAXONOMIC-LOCATION TRANSPARENCY

*is\_a* should be built in: **Ideally, each term's definition will represent the location in a term hierarchy to which that term belongs.** The principle of term-definition/taxonomic-location transparency essentially summarizes and requires that ontology construction abide by the principles that have already been discussed. If the principle of modularity and the Aristotelian top-down approach to term definition have been adhered to strictly, then the principle of term-definition/taxonomic location transparency will also be satisfied. Alternatively, violation of the principle of term-definition/taxonomic location transparency by a system of defined terms suggests that the principle of modularity or the Aristotelian approach or both have been violated. An ontology that adheres to this principle will be humanly intelligible and, as a general rule, computationally tractable insofar as all of the terms defined in the ontology will also stand in clear relationships to one another.<sup>19</sup>

#### CONSTRUCTION OF A TAXONOMIC HIERARCHY

Considered literally, a taxonomy is a tree-like structure consisting of nodes and branches, usually with a root node, leaf nodes, and intermediate nodes connected to each other, and to the root and leaf nodes by branches. Taxonomies are normally used to represent the hierarchical relationships amongst defined classes or universals in terms of the *is\_a* relationship.

Importantly, taxonomic structures can be generated amongst universals and defined classes in terms of a number of different relationships. For example, the *part\_of* relationship can be used to generate a taxonomic structure amongst universals. In such a case, the taxonomy generated might better be

---

<sup>18</sup> Barry Smith, Jacob Köhler and Anand Kumar. "[On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology](#)" in *Data Integration in the Life Sciences, First International Workshop, DILS 2004*, Leipzig, Germany, March 25-26, 2004, Proceedings, Editor: Erhard Rahm.

<sup>19</sup> Barry Smith, Jacob Köhler and Anand Kumar. "[On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology](#)" in *Data Integration in the Life Sciences, First International Workshop, DILS 2004*, Leipzig, Germany, March 25-26, 2004, Proceedings, Editor: Erhard Rahm.

referred to as a parontology. However, a parontology is not a taxonomy based on the *is\_a* relation. *is\_a* and *part\_of* have different meanings, and therefore must be kept strictly separate during the process of ontology design. Similarly, the relationship *descended\_from* can be used to generate taxonomic structures amongst biological species universals, as in phylogenetic trees.<sup>20</sup> Once again, such a *descends\_from* relationship is different from either *is\_a* or *part\_of*, and taxonomies based on these various relationships should be kept rigorously separate.

In what follows we will focus exclusively on taxonomies amongst universals structured by the *is\_a* relation.<sup>21</sup> This discussion thus focuses on taxonomies understood as consisting of “finitely many universals arranged in a tree-like hierarchy.”<sup>22</sup>

## CLASSIFICATION

Taxonomy is closely related to the issue of classifying entities. Indeed, a taxonomy just is one of the most common kinds of classifications of entities. But what is a classification? Relative to ontology, there are two major senses of the term ‘classification’.

The first has to do with identifying entities as instances of a given kind. This can happen in two ways. Particulars can be identified as instances of universals that they instantiate (as in “this cat is an instance of the kind cat”), and universals themselves can be identified as belonging to formal ontological categories (as in “the universal cat belongs in the formal ontological category of object or substance”). In this sense of ‘classification’, classifying an entity just involves recognizing what type of entity it is, either at the domain level, or at the level of formal ontological categories.

The second sense of ‘classification’ is the one that is directly related to taxonomies, though it always presupposes that some amount of classification in the first sense of the term has already taken place. In this second sense, a classification is a systematic organization of entities belonging to a given ontological category based on the relationships that these entities stand in both within and across ontological categories. To make this characterization concrete, consider the universals “eukaryotic cell” and “cell”. It is clear based on current biological knowledge that a eukaryotic cell *is\_a* cell. However, fully understanding the import of this *is\_a* relation rests on two things. First, the universals eukaryotic cell and cell are both substantial universals; instances of both universals are entities that persist as identical through time, gain and lose qualities and parts, and are wholly present at any time at which they exist at all. However, the universal eukaryotic cell is differentiated from the universal cell in virtue of its having as part a cell nucleus, such that eukaryotic cell *has\_part* cell nucleus is true.

---

<sup>20</sup> We did not define the *descends\_from* relation or anything like it in the earlier section on ontological categories and relations; however this could be done using the same basic strategy of starting with a primitive instance-level relation, and then defining the relationship amongst universals in terms of it.

<sup>21</sup> Much of the following discussion can be usefully compared with the more technically oriented proposals of Rector, A. L. “Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL.” <http://www.w3.org/TR/webont-reg>.

<sup>22</sup> Neuhaus, F., Grenon, P. & Smith, B. “A Formal Theory of Substances, Qualities and Universals”, in Achille Varzi and Laure Vieu (eds.), *Proceedings of FOIS 2004. International Conference on Formal Ontology and Information Systems*, Turin, 4–6 November 2004.

Similarly, plant cell *is\_a* eukaryotic cell. Both plant cell and eukaryotic cell are substantial universals, but plant cell is differentiated from eukaryotic cell by the possession of a cell wall as part, thus plant cell *has\_part* cell wall is true. So, what is going on here is that substantial universals (cell, eukaryotic cell, plant cell) are being classified in terms of their standing in a specific ontological relationship (*has\_part*) to other substantial universals (cell nucleus, cell wall). Notice that this also fits the Aristotelian definitional structure perfectly, “a eukaryotic cell (A) is a cell (B) that has a nucleus (C)”.

Similarly, the Aristotelian definition of human, “a human is an animal that is rational” classifies substantial universals into an *is\_a* hierarchy based on their standing in the *inheres\_in* relation to universals belonging to the ontological category of qualities. Thus rationality *inheres\_in* human, and differentiates the universal human from other substantial universals for kinds of animals.

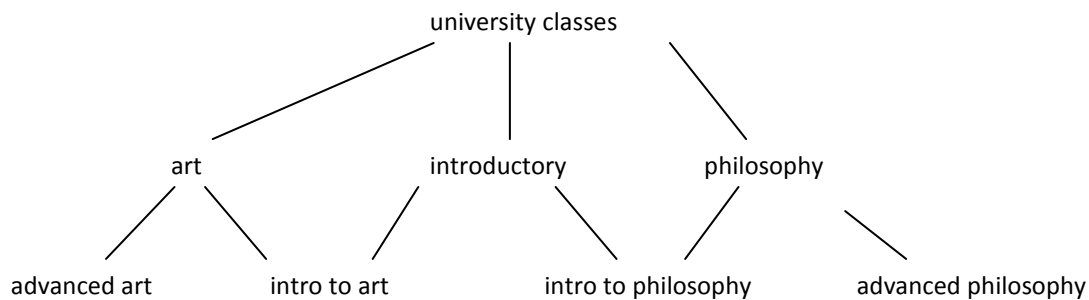
So, in this second (primary) sense, a classification is a systematic organization of entities belonging to a given ontological category based on the relationships that these entities stand in both within and across ontological categories. It is taxonomies in this sense that the universals in a good ontology should be organized in terms of. However, there is one more important feature of such classifications, and this is the principle in terms of which the classification has been generated.

#### THE PRINCIPLE OF CLASSIFICATION

Every classification should be structured according to the semantics for the *is\_a* relation and should identify distinct universals or classes only when these can be distinguished both from the universals or classes at the next level up in the hierarchy, and from other universals or classes at the same level, by some ontological feature or relation (such as possession of a characteristic part or quality). However, in addition to these features, classifications that form *is\_a* hierarchies are also normally structured in terms of a guiding principle or criterion of classification, one that further determines the specific meaning of the location of universals in the hierarchy, and determines, in part, which universals are relevant for inclusion in the hierarchy.

For example, it is possible to classify the classes in a university in terms of their subject matter, i.e. history, chemistry, biology, philosophy, etc. Alternatively, it is possible to classify the classes in a university in terms of their difficulty level, i.e. introductory, intermediate, advanced, graduate, etc. Each of these ‘ways of classifying’ university classes amounts to adopting a principle for the classification of university classes. **What is important is that, for any given classification, the principle that is being used be specified as clearly as possible at the outset, and then consistently adhered to throughout.** Further, two different principles should not be applied at the same level in the same hierarchy. Thus, a classification that attempted to simultaneously classify university classes both by subject matter and by difficulty level would end up looking something like the following:

#### Diagram 1.



The problem with this hierarchy is that saying *A is\_a B* is ambiguous. For example, 'art *is\_a* university class' can mean either that art is a university class of a particular kind *or* that art is a university class of a specific difficulty level. In other words, the relation '*is\_a*' in this hierarchy is ambiguous. And while it may be clear from context in this particular case which meaning the '*is\_a*' relation should be given in each particular context, in a more complicated case such as biological or medical science, such ambiguities, especially if they are perpetrated through an entire hierarchy of '*is\_a*' relations, are likely to lead to a great deal of confusion, at least for human users, and often also for automated reasoning.

More sophisticated principles of classification include similarity and difference of anatomical structure (one principle) or similarity and difference of genetic code (a second distinct principle) for organisms, atomic number for the elements in the periodic table, and kind of patient treated by (one principle) or kind of procedure performed by (a second distinct principle) for the classification of doctors in a hospital. Once more, what is crucial is that the principle being appealed to in classifying the entities in an *is\_a* hierarchy be both explicitly identified and consistently adhered to from the beginning.

**Up until this point the principles for best ontology practice that have been discussed have been intended to apply to all kinds of entities and domains whatsoever. The following principles for best ontology practice, though arguably having wider application as well, are primarily offered in the spirit of these traditional treatments of classification of substantial or objectual entities in terms of their characteristic qualities and or parts.<sup>23</sup>**

#### USE SINGLE INHERITANCE

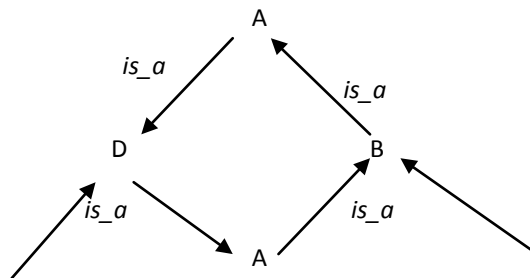
No diamonds: Based on the forgoing discussion, **an important principle for the classification of universals and classes in an ontology is to use single, not double- or multiple-, inheritance.** In a classification the relationship of inheritance is the relationship that a less abstract class stands in to the more abstract class that is directly above it in a classificatory hierarchy. Thus cat stands in the inheritance relationship to mammal, plant cell to eukaryotic cell, and eukaryotic cell to cell. Saying that a classification should use single inheritance means saying that every universal or class included in the classification should stand in

---

<sup>23</sup> See Barry Smith. "The Logic of Biological Classification and the Foundations of Biomedical Ontology." In Dag Westerstaahl (ed.), *Invited Papers from the 10<sup>th</sup> International Conference in Logic Methodology and Philosophy of Science*, Oviedo, Spain, 2003. Elsevier-North-Holland, 2004; Neuhaus, F., Grenon, P. & Smith, B. "A Formal Theory of Substances, Qualities and Universals", in Achille Varzi and Laure Vieu (eds.), *Proceedings of FOIS 2004. International Conference on Formal Ontology and Information Systems*, Turin, 4–6 November 2004.

an inheritance relationship to exactly one universal or class at the next highest level. When this principle is violated, taxonomies take on a diamond-like structure.

Such diamonds or 'multiple inheritance' should be avoided within a single classification in an ontology for a number of reasons. First, as they lead to an ambiguity in how the '*is\_a*' relation is to be interpreted in such classifications. Second, allowing such diamonds into a classificatory scheme can lead to the existence of loops within the classification, such that A *is\_a* B, B *is\_a* C, C *is\_a* D, and D *is\_a* A, as illustrated in diagram below. Allowing the existence of such loops in a classification amounts to adopting circular definitions for all or most of the terms located in the loop, and can lead to both human confusion and computational errors such as infinite loops.



The following is an example of a loop that can be found in the UMLS:

“Topographic regions: General terms

Physical anatomical entity

Anatomical spatial entity

Anatomical surface

Body regions

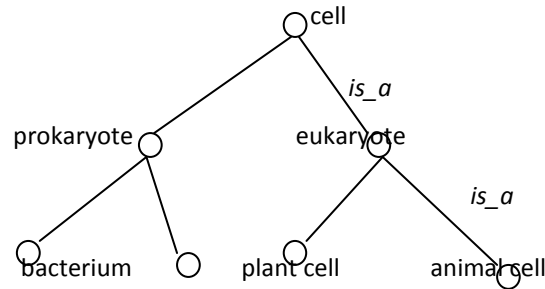
Topographic regions”<sup>24</sup>

A third problem with multiple inheritance is that it can lead to double-counting and hence to double-naming of entities in a classification. For example, the class “Intro to Art” from diagrams 7 inherits its properties from both the class “Introductory classes” and the class “Art classes”; if care is not taken, it can easily happen that a class inheriting from multiple super-classes will be taken as itself identifying multiple classes (double or multiple counting), each of which requires a separate name (hence double or multiple naming).

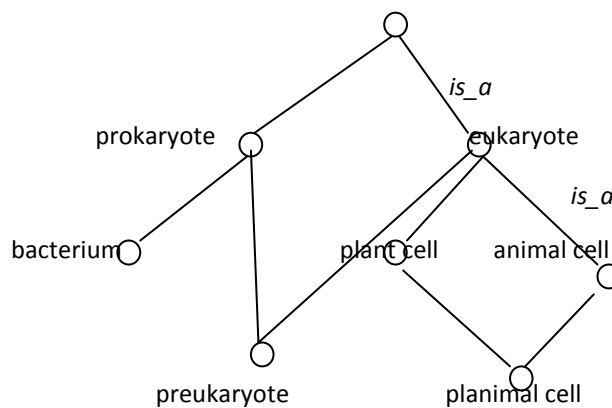
---

<sup>24</sup> See Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. Proc AMIA Symp 2001:57-61.

A fourth and final problem with allowing multiple inheritance into an ontology is the following: Each classification should involve one and only one principle of classification in order to avoid the possibility of ambiguity in its interpretation, as has already been discussed. Ideally, the principles used for organizing universals in a scientific classification will be as closely related as possible to the essential features of the entities that instantiate those universals. Thus good candidate principles for the classification of biological species include similarity and difference of anatomical structure, on the one hand, and similarity and difference of genetic material on the other. It is largely for biological scientists to decide which of these features is more essential in the identification and classification of biological species. However, what is important here is that if the classification of universals in a domain is carried out by specifying essential features and, if the Aristotelian procedure and definitional structure (an A (term) is a B (genus) that Cs (differential)) is adhered to, then universals at each level should be identified specifically in terms of features that *distinguish* them from all other universals or classes at that level. This means that no two universals at the same level should have any instances in common, and in turn means that no sub-universals or sub-classes of these universals should have any instances in common. Consider the classification of kinds of cells in **Diagram 2**.



The classification in Diagram 2 is consistent with the following Aristotelian structured definitions: “Prokaryote is a cell that lacks a nucleus”, “Eukaryote is a cell in which the genetic material is organized into a membrane-bound nucleus”, “Plant cell is a eukaryote cell that has a large central vacuole and a cell wall”, and “Animal cell is a eukaryote cell that has a small central vacuole and lacks a cell wall.” Now, given these definitions, suppose we were to allow the two instances of multiple inheritance from **Diagram 3** to occur.



The Aristotelian definition of ‘preukaryote’ would be “a prokaryote and eukaryote that has genetic material that is organized into a membrane-bound nucleus and lacks a nucleus”, while a ‘planimal cell’



would be ‘a eukaryote cell that has a large central vacuole and a cell wall, and has a small central vacuole and lacks a cell wall’. Now, the classes ‘preukaryote’ and ‘planimal cell’ should strike us as absurd. Their definitions contain manifest contradictions, and they seem like totally arbitrary and contrived denizens of the domain of cells. Note, however, that these classes were derived simply by allowing multiple inheritance into a classification that was well structured and organized according to a single principle along the lines of essential features of the universals in its domain, namely the kinds of cells and their essential differences in terms of possessing or not possessing certain kinds of parts (cell walls, a nucleus, etc.). Allowing multiple inheritance into a good classification leads to manifest absurdity, while a classification that straightforwardly yields multiple inheritance is probably a bad classification to begin with, either using a non-essential principle of classification for its classes and universals, or ambiguously attempting to apply multiple principles at the same time. While there are certainly cases of classification that are more complicated than cell biology, and where the choice of a principle of classification is not entirely clear, these cases are not different in kind, but only in degree of complexity from the case here discussed.<sup>25</sup>

**Note:** Importantly, it is often possible to classify a single universal in more than one way. For example, doctors can be classified in terms of the kinds of patients that they treat on the one hand, or in terms of the kinds of procedures that they perform on the other, and a term such as ‘pediatric surgeon’ could be classified in both of these ways (and would probably occupy a rather different place in the two classifications). However, in such cases, the answer is not to allow one taxonomy with multiple inheritance, but rather to construct two separate classifications, and use the definitions of the terms that appear in them, as well as the formal ontology (categories and relations) serving as the background of the domain ontology, to spell out the important relations between these two (or more) separate and diamondless classifications.<sup>26</sup>

#### JOINT EXHAUSTIVENESS

Don’t leave relevant universals out: **When classifying kinds of entities in a given domain, as much care as possible should be taken to ensure that all relevant universals are included at each level in a taxonomy.** An ideal classification would include all existing domain universals along with identifying and differentiating information for each, at each level in the hierarchy of organization. This does not mean that the designers of an ontology should sit around waiting for new scientific information (of which there is always more) before completing their ontology and making it available for use, but it does mean that all relevant domain universals that are discussed in contemporary domain literature and by contemporary domain experts should be included.

---

<sup>25</sup> Barry Smith, Jacob Köhler and Anand Kumar. “[On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology](#)” in *Data Integration in the Life Sciences, First International Workshop, DILS 2004, Leipzig, Germany, March 25-26, 2004, Proceedings*, Editor: Erhard Rahm; Barry Smith. “[The Logic of Biological Classification](#) and the Foundations of Biomedical Ontology.” In Dag Westerstahl (ed.), *Invited Papers from the 10<sup>th</sup> International Conference in Logic Methodology and Philosophy of Science*, Oviedo, Spain, 2003. Elsevier-North-Holland, 2004.

<sup>26</sup> Bittner, Smith, “Normalizing Medical Ontologies using Basic Formal Ontology”, in *Kooperative Versorgung, Vernetzte Forschung, Ubiquitäre Information* (Proceedings of GMDS Innsbruck, 26–30 September 2004), Niebüll: Videel OHG, 199–201. <http://ontology.buffalo.edu/medo/gmds2004Norm.pdf#search=%22Bittner%2C%20Smith%2C%20%E2%80%9CNormalizing%20Medical%20Ontologies%20using%20Basic%20Formal%20Ontology%22>; see also Rector, A. L. “Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL.” <http://www.w3.org/TR/webont-req>.

## MUTUAL EXCLUSIVITY

No shared sub-classes: No two universals or classes in a classification should have any sub-classes in common. In keeping with the above example, in a classification in which eukaryotic cell and prokaryotic cell occur at the same level as separate universals, they should have no sub-classes or sub-universals (no multiple inheritance) in common. The requirement that the universals or classes at each level of a classification be mutually exclusive is a straightforward consequence of the prohibition of multiple inheritance, and conversely.

## CLASS POSITIVITY

Complements of classes are not themselves classes: The complement of a class is the class containing all of the entities that do not belong in that class. Thus the complement of the class “dog” is the class “non-dog”. As a general rule of thumb, class-complements should be avoided when selecting the classes for and constructing the classification hierarchies in an ontology. The only thing that all of the members of a class-complement are guaranteed to have in common is the fact that there is some other class to which they all do not belong. Thus saying, of a given sub-class that it is subsumed by, for example, the class of all “non-conifer trees” is providing very little information about that sub-class. Further, class-complements rarely pick out genuine universals on the side of reality. Thus including many class complements in one’s ontology is likely to render it less accurate to the facts of reality, and so less useful for both scientific and practical purposes.

There are exceptions to this rule, including some of the examples given above. For example, prokaryotic cells are distinguished from eukaryotic and from all other cells precisely by the fact that they lack a cell nucleus. This is, in effect, negative information used to define a class. However, in this particular case there is overwhelming scientific evidence to the effect that, at this level of generality, dividing up cell universals in this way does lead to a principled and exhaustive classification of all kinds of cells. In such cases including negative information, or even featuring it in the definition of a term and the demarcation of a class may be unavoidable. However, even in these cases every effort should be made to include some positive information about the kinds of entities being defined and classified as well.

## CLASS OBJECTIVITY

Which classes exist is not a function of the current state of biological knowledge. Genuine classes, that is, the universals treated by natural science in any given domain, are discovered, not invented or created. This fact suggests a certain kind of general attitude or mind-set that should be taken towards the identification of classes in an ontology, namely, one that seriously takes into account the best available scientific information about reality in any given domain, and attempts to systematically organize that information according to its most essential characteristics.<sup>27</sup>

## CLASS-UNIVOCITY

As with terms, so with classes: No distinctions without differences. Every class should be clearly distinct from every other class in the ontology with respect to at least one property or characteristic. The best that having two classes characterized by exactly the same set of properties can achieve is redundancy.

---

<sup>27</sup> Barry Smith, Jacob Köhler and Anand Kumar. [“On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology”](#) in *Data Integration in the Life Sciences, First International Workshop, DILS 2004, Leipzig, Germany, March 25-26, 2004, Proceedings*, Editor: Erhard Rahm.

Failure to adhere to the principle of class-univocity can also be problematic insofar as assigning different names to classes that are the same will be likely to lead to human and computational confusion.

#### HIERARCHICAL STRUCTURE

Classifications have a hierarchical structure: Given the discussion that has gone before, the point should already be quite clear, that the terms in a classificatory hierarchy should be divided into predetermined levels. A classification hierarchy should include terms both for more specific and for more general universals. Thus besides including “red”, “human” or “H<sub>2</sub>O” a good classification should also include, as separate entries, “color”, “animal” or “compound”. Explicitly including this information will generate a division of terms into levels of generality, as can be seen in the examples discussed above. Hierarchically structured classifications are also a straightforward result of strictly adhering to the Aristotelian structure and method for good definitions, of identifying entities in terms of essential characteristics, and of classifying entities consistently in terms of a single principle of classification.

#### SUMMARY

The process of regimenting the domain information for an ontology thus includes the following steps, all to be carried out in terms of the principles that have been put forward above. First, select the exact terms and the format of these terms that are to be included in the ontology, based on domain information that has already been gathered. Second, provide clear, scientifically accurate and logically coherent definitions for each of these terms. Third, explicitly recognize the place or places of each of these terms in a hierarchical classification of the domain information. When this task has been consistently carried out, the domain information should be ready for the last great step in ontology development: formalization and computer implementation.

#### ORGANIZING THE RELATIONSHIPS

Relations in  
biomedical ontologies

Participants are requested to read the included published paper, *Relations in Biomedical Ontologies*, for details on this topic.

## WRAP UP QUESTIONS AND FURTHER READING

**Goal:** The final section will be devoted to clarifying doubts, providing suggestions on particular use-cases the participants might have as well as identifying potential collaborators amongst the participants themselves.

## REFERENCES

1. Hey, T. and A.E. Trefethen, *Cyberinfrastructure for e-Science*. Science, 2005. **308**(5723): p. 817-21.
2. Fedoroff, N., S.A. Racunas, and J. Shrager, *Making Biological Computing Smarter*, in *The Scientist*. 2005. p. 20-21.
3. Cimino, J.J. and X. Zhu, *The practical impact of ontologies on biomedical informatics*. Methods Inf Med, 2006. **45 Suppl 1**: p. 124-35.
4. Yu, A.C., *Methods in biomedical ontology*. J Biomed Inform, 2006. **39**(3): p. 252-66.
5. Bodenreider, O. and R. Stevens, *Bio-ontologies: current trends and future directions*. Brief Bioinform, 2006. **7**(3): p. 256-74.
6. Ashburner, M., et al., *Gene ontology: tool for the unification of biology*. The Gene Ontology Consortium. Nat Genet, 2000. **25**(1): p. 25-9.
7. Khatri, P. and S. Draghici, *Ontological analysis of gene expression data: current tools, limitations, and open problems*. Bioinformatics, 2005. **21**(18): p. 3587-95.
8. GO Slim. [Web page] 2003, [cited 2003; Available from: <http://www.geneontology.org/GO.slims.shtml>].
9. Alterovitz, G., et al., *GO PaD: the Gene Ontology Partition Database*. Nucleic Acids Res, 2007. **35**(Database issue): p. D322-7.
10. Lowe, H.J. and G.O. Barnett, *Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches*. Jama, 1994. **271**(14): p. 1103-8.
11. Bodenreider, O., *Using UMLS semantics for classification purposes*. Proc AMIA Symp, 2000: p. 86-90.
12. Hersh, W., et al., *Evaluation of SAPHIRE: an automated approach to indexing and retrieving medical literature*. Proc Annu Symp Comput Appl Med Care, 1991: p. 808-12.
13. Rubin, D.L., et al., *A statistical approach to scanning the biomedical literature for pharmacogenetics knowledge*. J Am Med Inform Assoc, 2005. **12**(2): p. 121-9.
14. Suomela, B.P. and M.A. Andrade, *Ranking the whole MEDLINE database according to a large training set using text indexing*. BMC Bioinformatics, 2005. **6**: p. 75.
15. Hartel, F.W., et al., *Modeling a description logic vocabulary for cancer research*. J Biomed Inform, 2005. **38**(2): p. 114-29.
16. Sioutos, N., et al., *NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information*. J Biomed Inform, 2007. **40**(1): p. 30-43.
17. Shah, N., et al., *Ontology-based Annotation and Query of Tissue Microarray Data*. AMIA Annu Symp Proc, 2006: p. 709-13.
18. Bodenreider, O., *The Unified Medical Language System (UMLS): integrating biomedical terminology*. Nucleic Acids Res, 2004. **32**(Database issue): p. D267-70.
19. Cimino, J.J., *Review paper: coding systems in health care*. Methods Inf Med, 1996. **35**(4-5): p. 273-84.
20. Langlotz, C.P., *RadLex: a new method for indexing online educational materials*. Radiographics, 2006. **26**(6): p. 1595-7.
21. Martone, M.E., A. Gupta, and M.H. Ellisman, *E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains*. Nat Neurosci, 2004. **7**(5): p. 467-72.

22. Swedlow, J.R., et al., *Informatics and quantitative analysis in biological imaging*. Science, 2003. **300**(5616): p. 100-2.
23. Rosse, C. and J.L. Mejino, Jr., *A reference ontology for biomedical informatics: the Foundational Model of Anatomy*. J Biomed Inform, 2003. **36**(6): p. 478-500.
24. Rubin, D.L., et al., *Using ontologies linked with geometric models to reason about penetrating injuries*. Artif Intell Med, 2006. **37**(3): p. 167-76.
25. Brinkley, J.F., *Structural informatics and its applications in medicine and biology*. Acad Med, 1991. **66**(10): p. 589-91.
26. Rosse, C., et al., *Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base*. J Am Med Inform Assoc, 1998. **5**(1): p. 17-40.
27. Dameron, O., et al., *Towards a sharable numeric and symbolic knowledge base on cerebral cortex anatomy: lessons learned from a prototype*. Proc AMIA Symp, 2002: p. 185-9.
28. Kahn, C.E., Jr., D.S. Channin, and D.L. Rubin, *An ontology for PACS integration*. J Digit Imaging, 2006. **19**(4): p. 316-27.
29. Noy, N.F., et al., *Protege-2000: an open-source ontology-development and knowledge-acquisition environment*. AMIA Annu Symp Proc, 2003: p. 953.
30. Horrocks, I., *An ontology language for the semantic Web*. Ieee Intelligent Systems, 2002. **17**(2): p. 74-75.
31. Ball, C.A. and A. Brazma, *MGED standards: work in progress*. Omics, 2006. **10**(2): p. 138-44.
32. BioPax-Consortium. *BioPAX: Biological Pathways Exchange*. 2006 [cited 2006 Dec 2006]; Available from: <http://www.biopax.org/>.
33. Kanehisa, M., et al., *From genomics to chemical genomics: new developments in KEGG*. Nucleic Acids Res, 2006. **34**(Database issue): p. D354-7.
34. Karp, P.D., et al., *Expansion of the BioCyc collection of pathway/genome databases to 160 genomes*. Nucleic Acids Res, 2005. **33**(19): p. 6083-9.
35. Joshi-Tope, G., et al., *Reactome: a knowledgebase of biological pathways*. Nucleic Acids Res, 2005. **33 Database Issue**: p. D428-32.
36. Demir, E., et al., *PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways*. Bioinformatics, 2002. **18**(7): p. 996-1003.
37. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks*. Genome Res, 2003. **13**(11): p. 2498-504.
38. Stevens, R., et al., *TAMBIS: transparent access to multiple bioinformatics information sources*. Bioinformatics, 2000. **16**(2): p. 184-5.
39. Gupta, A., et al., *Towards a formalization of disease-specific ontologies for neuroinformatics*. Neural Netw, 2003. **16**(9): p. 1277-92.
40. Li, C., et al., *OntoQuest: exploring ontological data made easy*, in *Proceedings of the 32nd international conference on Very large data bases - Volume 32*. 2006, VLDB Endowment: Seoul, Korea.
41. Kuchinsky, A., et al. *Biological Storytelling: a software tool for biological information organization based upon narrative structure*. in *Advanced Visual Interfaces*. 2002. Trento, Italy.
42. Karp, P.D., *Pathway databases: a case study in computational symbolic theories*. Science, 2001. **293**(5537): p. 2040-4.
43. Gifford, D.K., *Blazing pathways through genetic mountains*. Science, 2001. **293**(5537): p. 2049-51.
44. Racunas, S.A., et al., *HyBrow: a prototype system for computer-aided hypothesis evaluation*. Bioinformatics, 2004. **20**(suppl\_1): p. i257-264.
45. Racunas, S.A., N. Shah, and N.V. Fedoroff. *A Contradiction-Based Framework for Testing Gene Regulation Hypotheses*. in *IEEE Bioinformatics*. 2003. Stanford University, Palo Alto, California: IEEE Computer Society.
46. Proteome. *Yeast Proteome Database*. 2001 [cited 2002 4/14/2002]; Available from: <http://www.proteome.com/YPDhome.html>.
47. Dameron, O., M.A. Musen, and B. Gibaud, *Using semantic dependencies for consistency management of an ontology of brain-cortex anatomy*. Artif Intell Med, 2007. **39**(3): p. 217-25.

48. Rubin, D.L., O. Dameron, and M.A. Musen, *Use of description logic classification to reason about consequences of penetrating injuries*. AMIA Annu Symp Proc, 2005: p. 649-53.
49. Boustil, A., et al. *Classification des compte-rendus mammographiques a partir d'une ontologie radiologique en OWL*. in *Extraction et gestion de Connaissances (EGC'2006)*. 2006.
50. Rzhetsky, A., et al., *A knowledge model for analysis and simulation of regulatory networks*. Bioinformatics, 2000. **16**(12): p. 1120-8.
51. Rubin, D.L., et al., *National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge*. Omics, 2006. **10**(2): p. 185-98.